No-regret Exploration in Contextual Reinforcement Learning

Aditya Modi
Computer Science and Engineering
University of Michigan

Ambuj Tewari
Department of Statistics
University of Michigan

Abstract

We consider the recently proposed reinforcement learning (RL) framework of Contextual Markov Decision Processes (CMDP), where the agent interacts with a (potentially adversarial) sequence of episodic tabular MDPs. In addition, a context vector determining the MDP parameters is available to the agent at the start of each episode, thereby allowing it to learn a context-dependent near-optimal policy. In this paper, we propose a no-regret online RL algorithm in the setting where the MDP parameters are obtained from the context using generalized linear mappings (GLMs). We propose and analyze optimistic and randomized exploration methods which make (time and space) efficient online updates. The GLM based model subsumes previous work in this area and also improves previous known bounds in the special case where the contextual mapping is linear. In addition, we demonstrate a generic template to derive confidence sets using an online learning oracle and give a lower bound for the setting.

1 INTRODUCTION

Recent advances in reinforcement learning (RL) methods have led to increased focus on finding practical RL applications. RL algorithms provide a set of tools for tackling sequential decision making problems with potential applications ranging from web advertising and portfolio optimization, to healthcare applications like adaptive drug treatment. However, despite the empirical success of RL in simulated domains such as boardgames and video games, it has seen limited use in real world applications because of the inherent trial-and-error nature of the paradigm. In addition to these concerns, for the applications listed above, we have to essentially design adaptive methods for a *population* of users instead of a single system. For instance, optimizing adaptive drug treatment plans for an influx of patients has two key requirements: (1) ensure quickly learning good policies for each user and (2) share the observed outcome data efficiently across patients. Intuitively, we expect that frequently seen patient types (with some notion of similarity) can be adequately dealt with by using adaptive learning methods whereas difficult and rare cases could be carefully referred to experts to safely generate more data.

An efficient and plausible way to incorporate this heterogeneity is to include any distinguishing exogenous factors in form of a contextual information vector in the learning process. This information can include demographic, genomic features or individual measurements taken from lab tests. We model this setting using the framework of Contextual Markov Decision Processes (CMDPs) (Modi et al., 2018) where the learner has access to some *contextual features* at the start of every patient interaction. Similar settings have been studied with slight variations by Abbasi-Yadkori and Neu (2014); Hallak et al. (2015) and Dann et al. (2019). While the framework proposed in these works is innovative, there are a number of deficiencies in the available set of results. First, theoretical guarantees (PAC-style mistake bounds or regret bounds) sometimes hold only under a linearity assumption on the mapping between contexts and MDPs. This assumption is quite restrictive as it enforces additional constraints on the context features which are harder to satisfy in practice. Second, if non-linear mappings are introduced (Abbasi-Yadkori and Neu, 2014), the next state distributions are left un-normalized and therefore do not correctly model the context dependence of MDP dynamics.

We address these deficiencies by considering generalized linear models (GLMs) for mapping context features to MDP parameters (succinctly referred to as GLM-CMDP). We build upon the existing work on generalized linear bandits (Zhang et al., 2016) and propose UCRL2 (optimistic) and RLSVI (randomized) like algorithms with

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

regret analyses. Overall, our contributions are as follows:

- We provide optimistic and randomized regret minimizing algorithms for GLM-CMDPs. Our model subsumes/corrects previous CMDP frameworks and our analysis improves on the existing regret bounds by a factor of $\mathcal{O}(\sqrt{S})$ in the linear case.
- The proposed algorithms use *efficient online updates*, both in terms of memory and time complexity, improving over typical OFU approaches whose running time scales linearly with number of rounds.
- We prove a regret lower bound for GLM-CMDP when a logistic or quadratic link function is used.
- We provide a generic way to convert any online no-regret algorithm for estimating GLM parameters to confidence sets. This allows an improvement in the regret incurred by our methods when the GLM parameters have additional structure (e.g., sparsity).

2 SETTING AND NOTATION

We consider episodic Markov decision processes, denoted by tuple $(\mathcal{S}, \mathcal{A}, P, R, H)$ where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces, $P(\cdot|s,a)$ the transition distribution, $R(s,a)$ the reward function with mean $r(s,a)$ and H is the horizon. Without loss of generality, we will consider a fixed start state for each episode. In the contextual MDP setting (Hallak et al., 2015; Modi et al., 2018), the agent interacts with a sequence of MDPs M_k (indexed by k) whose dynamics and reward functions (denoted by P_k and R_k) are determined by an observed context vector $x_k \in \mathcal{X}$. For notation, we use $(s_{k,h}, a_{k,h}, r_{k,h}, s_{k,h+1})$ to denote the transition at step h in episode k. We denote the size of MDP parameters by the usual notation: $|\mathcal{S}| = S$ and $|\mathcal{A}| = A$.

The value of a policy in an episode k is defined as the expected total return for H steps in MDP M_k:

$$v_k^\pi = \mathbb{E}_{M_k,\pi}\Big[\sum_{h=1}^{H} r_{k,h}\Big]$$

The optimal policy for episode k is denoted by $\pi_k^* := \arg\max_\pi v_k^\pi$ and its value as v_k^*. The agent's goal in the CMDP setting is to learn a context dependent policy $\pi : \mathcal{X} \times \mathcal{S} \to \mathcal{A}$ such that cumulative expected return over K episodes is maximized. We quantify the agent's performance by the total regret incurred over a (potentially adversarial) sequence of K contexts:

$$R(K) := \sum_{k=1}^{K} v_k^* - v_k^{\pi_k} \tag{1}$$

Note that the regret here is defined with respect to the sequence of context dependent optimal policies.

Additional notation. For two matrices X and Y, the inner product is defined as $\langle X, Y\rangle := \mathrm{Tr}(X^\top Y)$. For a vector $x \in \mathbb{R}^d$ and a matrix $A \in \mathbb{R}^{d\times d}$, we define $\|x\|_A^2 := x^\top A x$. For matrices $W \in \mathbb{R}^{m\times n}$ and $X \in \mathbb{R}^{n\times n}$, we define $\|W\|_X^2 := \sum_{i=1}^m \|W^{(i)}\|_X^2$ where $W^{(i)}$ is the i^{th} row of the matrix. Further, we reserve the notation $\|W\|_F$ to denote the Frobenius norm of a matrix W. Any norm which appears without a subscript will denote the ℓ_2 norm for a vector and the Frobenius norm for a matrix.

2.1 GENERALIZED LINEAR MODEL FOR CMDPs

Using a linear mapping of the predictors is a simple and ubiquitous approach for modeling contextual/dynamical dependence in sequential decision making problems. Linear models are also well known for being interpretable and explainable, properties which are very valuable in our motivating settings. Similarly, we also utilize this structural simplicity of linearity and model the categorical output space $(p(\cdot|s,a))$ in a contextual MDP using generalized linear mappings. Specifically, for each pair $s, a \in \mathcal{S} \times \mathcal{A}$, there exists a weight matrix $W_{sa} \in \mathcal{W} \subseteq \mathbb{R}^{S\times d}$ where $\mathcal{W}$ is a convex set. For any context $x_k \in \mathbb{R}^d$, the next state distribution for the pair is specified by a GLM:

$$P_k(\cdot|s,a) = \nabla\Phi(W_{sa}x_k) \tag{2}$$

where $\Phi(\cdot) : \mathbb{R}^S \to \mathbb{R}$ is the link function of the GLM[1]. We will assume that this link function is convex which is always the case for a canonical exponential family (Lauritzen, 1996). For rewards, we assume that each mean reward is given by a linear function[2] of the context: $r_k(s,a) := \theta_{sa}^\top x_k$ where $\theta \in \Theta \subseteq \mathbb{R}^d$. In addition, we will make the following assumptions about the link function.

Assumption 2.1. *The function $\Phi(\cdot)$ is α-strongly convex and β-strongly smooth, that is:*

$$\Phi(v) \geq \Phi(u) + \langle\nabla\Phi(u), v-u\rangle + \tfrac{\alpha}{2}\|u-v\|_2^2 \tag{3}$$

$$\Phi(v) \leq \Phi(u) + \langle\nabla\Phi(u), v-u\rangle + \tfrac{\beta}{2}\|u-v\|_2^2 \tag{4}$$

We will see that this assumption is critical for constructing the confidence sets used in our algorithm. We make another assumption about the size of the weight matrices W_{sa}^* and contexts x_k:

[1] We abuse the term GLM here as we don't necessarily consider a complementary exponential family model in eq. (2)

[2] Similar results can be derived for GLM reward functions.

Assumption 2.2. *For all episodes k, we have $\|x_k\|_2 \leq R$ and for all state-action pairs (s,a), $\|W_{sa}^{(i)}\|_2 \leq B_p$ and $\|\theta_{sa}\|_2 \leq B_r$. So, we have $\|Wx_k\|_\infty \leq B_p R$ for all $W \in \mathcal{W}$.*

The following two contextual MDP models are special cases of our setting:

Example 2.3 (Multinomial logit model, Agarwal (2013))**.** *Each next state is sampled from a categorical distribution with probabilities*[3]*:*

$$P_x(s_i|s,a) = \frac{\exp(W_{sa}^{(i)} x)}{\sum_{j=1}^{S} \exp(W_{sa}^{(j)} x)}$$

The link function for this case can be given as $\Phi(y) = \log(\sum_{i=1}^{S} \exp(y_i))$ which can be shown to be strongly convex with $\alpha = \frac{1}{\exp(BR)S^2}$ and smooth with $\beta = 1$.

Example 2.4 (Linear combination of MDPs, Modi et al. (2018))**.** *Each MDP is obtained by a linear combination of d base MDPs $\{(\mathcal{S}, \mathcal{A}, P^i, R^i, H)\}_{i=1}^d$. Here, $x_k \in \Delta_{d-1}$*[4]*, and $P_k(\cdot|s,a) := \sum_{i=1}^d x_{ki} P^i(\cdot|s,a)$. The link function for this can be shown to be:*

$$\Phi(y) = \tfrac{1}{2}\|y\|_2^2$$

which is strongly convex and smooth with parameters $\alpha = \beta = 1$. Moreover, W_{sa} here is the $S \times d$ matrix containing each next state distribution in a column. We have, $B_p \leq \sqrt{d}$, $\|W_{sa}\|_F \leq \sqrt{d}$ and $\|W_{sa} x_k\|_2 \leq 1$.

3 ONLINE ESTIMATES AND CONFIDENCE SET CONSTRUCTION

In order to obtain a no-regret algorithm for our setting, we will follow the popular *optimism in the face of uncertainty* (OFU) approach which relies on the construction of confidence sets for MDP parameters at the beginning of each episode. We focus on deriving these confidence sets for the next state distributions for all state action pairs. We assume that the link function Φ and values α, B and R are known a priori. The confidence sets are constructed and used in the following manner in the OFU template for MDPs: at the beginning of each episode $k = 1, 2, \ldots, K$:

- For each (s,a), compute an estimate of transition distribution $\widehat{P}_k(\cdot|s,a)$ and mean reward $\hat{r}_k(s,a)$ along with confidence sets $\mathcal{P}$ and $\mathcal{R}$ such that $P_k(\cdot|s,a) \in \mathcal{P}$ and $r_k(s,a) \in \mathcal{R}$ with high probability.
- Compute an optimistic policy π_k using the confidence sets and unroll a trajectory in M_k with π_k. Using observed transitions, update the estimates and confidence sets.

Therefore, in the GLM-CMDP setup, estimating transition distributions and reward functions is the same as estimating the underlying parameters W_{sa} and θ_{sa} for each pair (s,a). Likewise, any confidence set $\mathcal{W}_{sa}$ for W_{sa} (Θ_{sa} for θ_{sa}) can be translated into a confidence set of transition distributions.

In our final algorithm for GLM-CMDP, we will use the method from this section for estimating the next state distribution for each state-action pair. The reward parameter θ_{sa} and confidence set Θ_{sa} is estimated using the linear bandit estimator (Lattimore and Szepesvri (2020), Chap. 20). Here, we solely focus on the following online estimation problem without any reference to the CMDP setup. Specifically, given a link function Φ, the learner observes a sequence of contexts $x_t \in \mathcal{X}$ ($t = 1, 2, \ldots$) and a sample y_t drawn from the distribution $P_t \equiv \nabla\Phi(W^* x_t)$ over a finite domain of size S. Here, we use W^* to denote the true parameter for the given GLM model. The learner's task is to compute an estimate W_t for W^* and a confidence set $\mathcal{W}_t$ after any such t samples. We frame this as an online optimization problem with the following loss sequence (based on the negative log-likelihood):

$$l_t(W; x_t, y_t) = \Phi(Wx_t) - y_t^\top W x_t \tag{5}$$

where y_t is the one-hot representation of the observed sample in round t. This loss function preserves the strong convexity of Φ with respect to Wx_t and is a proper loss function (Agarwal, 2013):

$$\arg\min_W \mathbb{E}\big[l_t(W; x_t, y_t)|x_t\big] = W^* \tag{6}$$

Since our aim is computational and memory efficiency, we carefully follow the Online Newton Step (Hazan et al., 2007) based method proposed for $0/1$ rewards with logistic link function in Zhang et al. (2016). While deriving the confidence set in this extension to GLMs, we use properties of categorical vectors in various places in the analysis which eventually saves a factor of S. The online update scheme is shown in Algorithm 1. Interestingly, note that for tabular MDPs, where $d = \alpha = 1$ and $\Phi(y) = \frac{1}{2}\|y\|_2^2$, with $\eta = 1$, we would recover the empirical average distribution as the online estimate. Along with the estimate W_{t+1}, we can also construct a high probability confidence set as follows:

Theorem 3.1 (Confidence set for W^*)**.** *In Algorithm 1, for all timesteps $t = 1, 2, \ldots$, with probability at least $1 - \delta$, we have:*

$$\|W_{t+1} - W^*\|_{Z_{t+1}} \leq \sqrt{\gamma_{t+1}} \tag{8}$$

[3] Without loss of generality, we can set the last row $W_{sa}^{(S)}$ of the weight matrix to be 0 to avoid an overparameterized system.

[4] Δ_{d-1} denotes the simplex $\{x \in \mathbb{R}^d : \|x\|_1 = 1,\ x \geq 0\}$.

Algorithm 1 Online parameter estimation for GLMs

Input: Φ, α, η
Set $W_1 \leftarrow \mathbf{0}$, $Z_1 \leftarrow \lambda \mathbb{I}_d$
for $t = 1, 2, \ldots$ **do**
  Observe x_t and sample $y_t \sim P_t(\cdot)$
  Compute new estimate W_{t+1}:

$$\arg\min_{W \in \mathcal{W}} \frac{\|W - W_t\|^2_{Z_{t+1}}}{2} + \eta \langle \nabla l_t(W_t x_t) x_t^\top, W - W_t \rangle \tag{7}$$

  where $Z_{t+1} = Z_t + \frac{\eta\alpha}{2} x_t x_t^\top$.

where

$$\gamma_{t+1} = \lambda B^2 + 8\eta B_p R + 2\eta \Big[(\tfrac{4}{\alpha} + \tfrac{8}{3} B_p R)\tau_t + \tfrac{4}{\alpha} \log \tfrac{\det(Z_{t+1})}{\det(Z_1)} \Big] \tag{9}$$

with $\tau_t = \log(2\lceil 2 \log St \rceil t^2/\delta)$ *and* $B = \max_{W \in \mathcal{W}} \|W\|_F$.

Any upper bound for $\|W^*\|_F^2$ can be substituted for B the confidence width in eq (9). The term γ_t depends on the size of the true weight matrix, strong convexity parameter $\frac{1}{\alpha}$ and the log determinant of the covariance matrix. We will later show that the last term grows at a $\mathcal{O}(d \log t)$ rate. Therefore, overall γ_t scales as $\mathcal{O}(S + \frac{d}{\alpha} \log^2 t)$. The complete proof can be found in Appendix A.

Algorithm 1 only stores the empirical covariance matrix and solves the optimization problem (7) for the current context. Since $\mathcal{W}$ is convex, this is a tractable problem and can be solved via any off-the-shelf optimizer up to desired accuracy. The total computation time for each context and all (s, a) pairs is $O(\text{poly}(S, A, d))$ with no dependence on t. Furthermore, we only store SA-many matrices of size $S \times d$ and covariance matrices of sizes $d \times d$. Thus, both time and memory complexity of the method are bounded by $\mathcal{O}(\text{poly}(S, A, H, d))$ per episode.

4 NO-REGRET ALGORITHMS FOR GLM-CMDP

4.1 OPTIMISTIC REINFORCEMENT LEARNING FOR GLM CMDP

In this section, we describe the OFU based online learning algorithm which leverages the confidence sets as described in the previous section. Not surprisingly, our algorithm is similar to the algorithm of Dann et al. (2019) and Abbasi-Yadkori and Neu (2014) and follows the standard format for no-regret bounds in MDPs. In all discussions about CMDPs, we will again use $x_k \in \mathcal{X}$ to denote the context for episode k and use Algorithm 1 from the previous section to estimate the corresponding MDP M_k. Specifically, for each state-action pair (s, a), we use all observed transitions to estimate W_{sa} and θ_{sa}. We compute and store the quantities used in Algorithm 1 for each (s, a): we use $\widehat{W}_{k,sa}$ to denote the parameter estimate for W_{sa} at the beginning of the k^{th} episode. Similarly, we use the notation $\gamma_{k,sa}$ and $Z_{k,sa}$ for the other terms. Using the estimate $\widehat{W}_{k,sa}$ and the confidence set, we compute the confidence interval for $P_k(\cdot|s, a)$:

$$\begin{aligned} \xi^{(p)}_{k,sa} &\coloneqq \|P_k(\cdot|s,a) - \widehat{P}_k(\cdot|s,a)\|_1 \\ &\le \beta\sqrt{S} \|W_{sa} - \widehat{W}_{k,sa}\|_{Z_{k,sa}} \|x_k\|_{Z^{-1}_{k,sa}} \\ &\le \beta\sqrt{S}\sqrt{\gamma_{k,sa}} \|x_k\|_{Z^{-1}_{k,sa}} \end{aligned} \tag{10}$$

where in the definition of $\gamma_{k,sa}$ we use $\delta = \delta_p$. It is again easy to see that for tabular MDPs with $d = 1$, we recover a similar confidence interval as used in Jaksch et al. (2010). For rewards, using the results from linear contextual bandit literature (Lattimore and Szepesvri (2020), Theorem 20.5), we use the following confidence interval:

$$\begin{aligned} \xi^{(r)}_{k,sa} &\coloneqq |r_k(s,a) - \hat{r}_k(s,a)| \\ &= \underbrace{\left(\sqrt{\lambda d} + \sqrt{\tfrac{1}{4} \log \tfrac{\det Z_{k,sa}}{\delta_r^2 \det \lambda I}} \right)}_{\coloneqq \zeta_{k,sa}} \|x_k\|_{Z^{-1}_{k,sa}} \end{aligned} \tag{11}$$

In GLM-ORL, we use these confidence intervals to compute an optimistic policy (Lines 9-15). The computed value function is optimistic as we add the total uncertainty as a bonus (Line 11) during each Bellman backup. For any step h, we clip the optimistic estimate between $[0, H - h]$ during Bellman backups (Line 13[5]). After unrolling an episode using π_k, we update the parameter estimates and confidence sets for every observed (s, a) pair.

For any sequence of K contexts, we can guarantee the following regret bound:

Theorem 4.1 (Regret of GLM-ORL)**.** *For any $\delta \in (0, 1)$, if Algorithm 2 is run with the estimation method 1, then for all $K \in \mathbb{N}$ and with probability at least $1 - \delta$, the regret $R(K)$ is:*

$$\tilde{\mathcal{O}}\Big(\Big(\frac{\sqrt{d} \max_{s,a} \|W_{sa}\|_F}{\sqrt{\alpha}} + \frac{d}{\alpha} \Big) \beta S H^2 \sqrt{AK} \log \frac{KHd}{\lambda\delta} \Big)$$

If $\|W^{(i)}\|$ is bounded by B_p, we get $\|W_{sa}\|_F^2 \le S B_p^2$, whereas, for the linear case (Ex. 2.4), $\|W_{sa}\|_F^2 \le \sqrt{d}$. Substituting the bounds on $\|W_{sa}\|_F^2$, we get:

Corollary 4.2 (Multinomial logit model)**.** *For example 2.3, we have $\|W\|_F \le B\sqrt{S}$, $\alpha = \frac{1}{\exp(BR) S^2}$ and $\beta = 1$. Therefore, the regret bound of Algorithm 2 is $\tilde{\mathcal{O}}(dS^3 H^2 \sqrt{AK})$.*

[5] We use the notation $a \wedge b$ to denote $\min(a, b)$ and $a \vee b$ for $\max(a, b)$.

Corollary 4.3 (Regret bound for linear combination case)**.** *For example 2.4, with $\|W\|_F \leq \sqrt{d}$, the regret bound of Algorithm 2 is $\tilde{\mathcal{O}}(dSH^2\sqrt{AK})$.*

Algorithm 2 GLM-ORL (GLM Optimistic Reinforcement Learning)

Input: $\mathcal{S}, \mathcal{A}, H, \Phi, d, \mathcal{W}, \lambda, \delta$
$\delta' = \frac{\delta}{2SA+SH}$, $\widetilde{V}_{k,H+1}(s) = 0 \; \forall s \in \mathcal{S}, k \in \mathbb{N}$
for $k \leftarrow 1, 2, 3, \ldots$ **do**
  Observe current context x_k
  for $s \in \mathcal{S}, a \in \mathcal{A}$ **do**
    $\widehat{P}_k(\cdot|s,a) \leftarrow \nabla\Phi(\widehat{W}_{k,sa}x_k)$
    $\hat{r}_k(s,a) \leftarrow \langle \hat{\theta}_{k,sa}, x_k \rangle$
    Compute conf. intervals using eqns. (10), (11)
  for $h \leftarrow H, H-1, \cdots, 1$, and $s \in \mathcal{S}$ **do**
    for $a \in \mathcal{A}$ **do**
      $\varphi = \|\widetilde{V}_{k,h+1}\|_\infty \xi^{(p)}_{k,sa} + \xi^{(r)}_{k,sa}$
      $\widetilde{Q}_{k,h}(s,a) = \widehat{P}_{k,sa}^\top \widetilde{V}_{k,h+1} + \hat{r}_k(s,a) + \varphi$
      $\widetilde{Q}_{k,h}(s,a) = 0 \vee (\widetilde{Q}_{k,h}(s,a) \wedge V_h^{\max})$
    $\pi_{k,h}(s) = \arg\max_a \widetilde{Q}_{k,h}(s,a)$
    $\widetilde{V}_{k,h}(s) = \widetilde{Q}_{k,h}(s, \pi_{k,h}(s))$
  Unroll a trajectory in M_k using π_k
  Update $\widehat{W}_{sa}$ and $\hat{\theta}_{sa}$ for observed samples.

In Corollary 4.3, the bound is worse by a factor of $\sqrt{H}$ when compared to the $\widetilde{\mathcal{O}}(HS\sqrt{AKH})$ bound of UCRL2 for tabular MDPs ($d = 1$). This factor is incurred while bounding the sum of confidence widths in eq. (28) (in UCRL2 it is $\mathcal{O}(\sqrt{SAKH})$).

4.1.1 Proof of Theorem 4.1

We provide the key lemmas used in the analysis with the complete proof in Appendix B.1. Here, we assume that transition probability estimates are valid with probability at least $1 - \delta_p$ and reward estimates with $1 - \delta_r$ for all (s, a) for all episodes. We first begin by showing that the computed policy's value is optimistic.

Lemma 4.4 (Optimism)**.** *If all the confidence intervals as computed in Algorithm 2 are valid for all episodes k, then for all k and $h \in [H]$ and $s, a \in \mathcal{S} \times \mathcal{A}$, we have:*

$$\tilde{Q}_{k,h}(s,a) \geq Q^*_{k,h}(s,a)$$

Proof. We show this via an inductive argument. For every episode, the lemma is true trivially for $H + 1$. Assume that it is true for $h + 1$. For h, we have:

$$\begin{aligned}
&\tilde{Q}_{k,h}(s,a) - Q^*_{k,h}(s,a) \\
&= (\widehat{P}_k(s,a)^\top \tilde{V}_{k,h+1} + \hat{r}_k(s,a) + \varphi_{k,h}(s,a)) \wedge V_h^{\max} \\
&\quad - P_k(s,a)^\top V^*_{k,h+1} - r_k(s,a)
\end{aligned}$$

We use the fact that when $\tilde{Q}_{k,h}(s,a) = V_h^{\max}$, the lemma is trivially satisfied. When $\tilde{Q}_{k,h}(s,a) < V_h^{\max}$, we have:

$$\begin{aligned}
&\tilde{Q}_{k,h}(s,a) - Q^*_{k,h}(s,a) \\
&= \hat{r}_k(s,a) - r_k(s,a) + \widehat{P}_k(s,a)^\top (\tilde{V}_{k,h+1} - V^*_{k,h+1}) \\
&\quad + \varphi_{k,h}(s,a) - (P_k(s,a) - \widehat{P}_k(s,a))^\top V^*_{k,h+1} \\
&\geq -|\hat{r}_k(s,a) - r_k(s,a)| + \varphi_{k,h}(s,a) \\
&\quad - \|P_k(s,a) - \widehat{P}_k(s,a)\|_1 \|\tilde{V}_{k,h+1}\|_\infty \geq 0
\end{aligned}$$

The last step uses the guarantee on confidence intervals and the inductive assumption for $h + 1$. Therefore, the estimated Q-values are optimistic by induction. □

Using this optimism guarantee, we can bound the instantaneous regret Δ_k in episode k as: $V^*_{k,1}(s) - V^{\pi_k}_{k,1}(s) \leq \widetilde{V}_{k,1}(s) - V^{\pi_k}_{k,1}(s)$. With $\widetilde{V}$ as the upper bound, we can bound the total regret with the following Lemma:

Lemma 4.5. *In the event that the confidence sets are valid for all episodes, then with probability at least $1 - SH\delta_1$, the total regret $R(K)$ can be bounded by*

$$\begin{aligned}
R(K) \leq SH\sqrt{K \log \tfrac{6 \log 2K}{\delta_1}} \\
+ \sum_{k=1}^{K} \sum_{h=1}^{H} \cdot (2\varphi_{k,h}(s_{k,h}, a_{k,h}) \wedge V_h^{\max})
\end{aligned} \tag{12}$$

The proof is given in the appendix. The second term in ineq. (12) can now be bounded as follows:

$$\begin{aligned}
&\sum_{k=1}^{K} \sum_{h=1}^{H} (2\varphi(s_{k,h}, a_{k,h}) \wedge V_h^{\max}) \\
&\leq \sum_{k=1}^{K} \sum_{h=1}^{H} (2\xi^{(r)}_{k,s_{k,h},a_{k,h}} \wedge V_h^{\max}) \\
&\quad + \sum_{k=1}^{K} \sum_{h=1}^{H} (2V_{h+1}^{\max} \xi^{(p)}_{k,s_{k,h},a_{k,h}} \wedge V_h^{\max})
\end{aligned} \tag{13}$$

We ignore the reward estimation error in eq. (13) as it leads to lower order terms. The second expression can be again bounded as follows:

$$\begin{aligned}
&\sum_{k=1}^{K} \sum_{h=1}^{H} (2V_{h+1}^{\max} \xi^{(p)}_{k,s_{k,h},a_{k,h}} \wedge V_h^{\max}) \\
&\leq 2 \sum_{k,h} V_h^{\max} \left(1 \wedge \beta \sqrt{S\gamma_k(s_{k,h}, a_{k,h})} \|x_k\|_{Z^{-1}_{k,sa,h}} \right)
\end{aligned} \tag{14}$$

Using Lemma B.4, we see that

$$\begin{aligned}\gamma_k(s,a) &\coloneqq f_\Phi(k,\delta_p) + \frac{8\eta}{\alpha}\log\frac{\det(Z_{k,sa})}{\det(Z_{1,sa})} \\ &\leq \frac{\eta\alpha}{2S} + f_\Phi(KH,\delta_p) + \frac{8\eta}{\alpha}\log\frac{\det(Z_{K+1,sa})}{\det(Z_{1,sa})} \\ &\leq \frac{\eta\alpha}{2S} + f_\Phi(KH,\delta_p) + \frac{8\eta d}{\alpha}\log\left(1+\frac{KHR^2}{\lambda d}\right)\end{aligned}$$

We use $f_\Phi(k,\delta_p)$ to refer to the Z_k independent terms in eq. (9). Setting $\bar{\gamma}_K$ to the last expression guarantees that $\frac{2S\bar{\gamma}_K}{\eta\alpha} \geq 1$. We can now bound the term in eq. (14) as:

$$\begin{aligned}&2\beta V_1^{\max}\sqrt{\frac{2S\bar{\gamma}_K}{\eta\alpha}}\sum_{k,h}\left(1\wedge\sqrt{\frac{\eta\alpha}{2}}\|x_k\|_{Z^{-1}_{k,sa,h}}\right) \\ &\leq 2\beta V_1^{\max}\sqrt{\frac{2S\bar{\gamma}_K KH}{\eta\alpha}}\sqrt{\sum_{k,h}\left(1\wedge\frac{\eta\alpha}{2}\|x_k\|^2_{Z^{-1}_{k,sa,h}}\right)} \qquad (15)\end{aligned}$$

Ineq. (15) follows by using Cauchy-Schwarz inequality. Finally, by using Lemma B.4 in Appendix B.1, we can bound the term as

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{h=1}^{H}(2V_{h+1}^{\max}\xi^{(p)}_{k,s_{k,h},a_{k,h}} \wedge V_h^{\max}) \\ &= 4\beta V_1^{\max}\sqrt{\frac{2S\bar{\gamma}_K KH}{\eta\alpha}}\sqrt{2HSAd\log\left(1+\frac{KHR^2}{\lambda d}\right)}\end{aligned}$$

Now, after setting the failure probabilities $\delta_1 = \delta_p = \delta_r = \delta/(2SA+SH)$ and taking a union bound over all events, we get the total failure probability as δ. Therefore, with probability at least $1-\delta$, we can bound the regret of GLM-ORL as

$$R(K) = \widetilde{\mathcal{O}}\left(\left(\frac{\sqrt{d}\max_{s,a}\|W^*_{sa}\|_F}{\sqrt{\alpha}} + \frac{d}{\alpha}\right)\beta SH^2\sqrt{AK}\right)$$

where $\max_{s,a}\|W^*_{sa}\|_F$ is replaced by the problem dependent upper bound assumed to be known a priori.[6]

4.1.2 Mistake bound for GLM-ORL

The regret analysis shows that the total value loss suffered by the agent is sublinear in K, and therefore, goes to 0 on average. However, this can still lead to infinitely many episodes where the sub-optimality gap is larger than a desired threshold ϵ, given that it occurs relatively infrequently. It is still desirable, for practical purposes, to analyze how frequently can the agent incur such mistakes. Here, a mistake is defined as an episode in which the value of the learner's policy π_k is not ϵ-optimal, i.e., $V_k^* - V_k^{\pi_k} \geq \epsilon$. In our setting, we can show the following result.

Theorem 4.6 (Bound on the number of mistakes)**.** *For any number of episodes K, $\delta \in (0,1)$ and $\epsilon \in (0,H)$, with probability at least $1-\delta$, the number of episodes where GLM-ORL's policy π_k is not ϵ-optimal is bounded by*

$$\mathcal{O}\left(\frac{dS^2AH^5\log(KH)}{\epsilon^2}\left(\frac{d\log^2(KH)}{\alpha}+S\right)\right)$$

ignoring $\mathcal{O}(poly(\log\log KH))$ terms.

We defer the proof to Appendix C. Note that this term depends poly-logarithmically on K and therefore increases with time. The algorithm doesn't need to know the value of ϵ and result holds for all ϵ. This differs from the standard mistake bound style PAC guarantees where a finite upper bound is given. Dann et al. (2019) argued that this is due to the non-shrinking nature of the constructed confidence sets. As such, showing such a result for CMDPs requires a non-trivial construction of confidence sets and falls beyond the scope of this paper.

4.2 RANDOMIZED EXPLORATION FOR GLM-CMDP

Empirical investigations in bandit and MDP literature has shown that optimism based exploration methods typically over-explore, often resulting in sub-optimal empirical performance. In contrast, Thompson sampling based methods which use randomization during exploration have been shown to have an empirical advantage with slightly worse regret guarantees. Recently, Russo (2019) showed that even with such randomized exploration methods, one can achieve a worst-case regret bound instead of the typical Bayesian regret guarantees. In this section, we show that the same is true for GLM-CMDP where a randomized reward bonus can be used for exploration. We build upon their work to propose an RLSVI style method (Algorithm 3) and analyze its expected regret. The main difference between Algorithm 2 and Algorithm 3 is that instead of the fixed bonus φ (Line 11) in the former, GLM-RLSVI samples a random reward bonus in Line 12 for each (s,a) from the distribution $N(0, HS\varphi^2)$. The variance term φ is set to a sufficiently high value, such that, the resulting policy is optimistic with constant probability. We use a slightly modified version of the confidence sets as follows:

$$\begin{aligned}\overline{\xi}^{(p)}_{k,sa} &\coloneqq 2\wedge\left(\beta\sqrt{S}\sqrt{\gamma_{k,sa}}\|x_k\|_{Z^{-1}_{k,sa}}\right) \\ \overline{\xi}^{(r)}_{k,sa} &\coloneqq B_r R\wedge\left(\tau_{k,sa}\|x_k\|_{Z^{-1}_{k,sa}}\right)\end{aligned}$$

[6] An improved dependence on $\sum_{s,a}\|W^*_{sa}\|_F$ can be obtained instead of $S\max_{s,a}\|W^*_{sa}\|_F$ in the regret bound.

Algorithm 3 GLM-RLSVI

Input: $\mathcal{S}, \mathcal{A}, H, \Phi, d, \mathcal{W}, \lambda$
$\overline{V}_{k,H+1}(s) = 0 \; \forall s \in \mathcal{S}, k \in \mathbb{N}$
for $k \leftarrow 1, 2, 3, \ldots$ **do**
  Observe current context x_k
  for $s \in \mathcal{S}, a \in \mathcal{A}$ **do**
    $\widehat{P}_k(\cdot|s,a) \leftarrow \nabla\Phi(\widehat{W}_{k,sa} x_k)$
    $\hat{r}_k(s,a) \leftarrow \langle \hat{\theta}_{k,sa}, x_k \rangle$
    Compute conf. intervals using eqns. (10), (11)
  for $h \leftarrow H, H-1, \cdots, 1$, and $s \in \mathcal{S}$ **do**
    for $a \in \mathcal{A}$ **do**
      $\varphi = (H-h)\overline{\xi}^{(p)}_{k,sa} + \overline{\xi}^{(r)}_{k,sa}$
      Draw sample $b_{k,h}(s,a) \sim N(0, SH\varphi)$
      $\overline{Q}_{k,h}(s,a) = \widehat{P}^{\top}_{k,sa}\overline{V}_{k,h+1} + \hat{r}_k(s,a) + b_{k,h}(s,a)$
    $\pi_{k,h}(s) = \arg\max_a \overline{Q}_{k,h}(s,a)$
    $\overline{V}_{k,h}(s) = \overline{Q}_{k,h}(s, \pi_{k,h}(s))$
  Unroll a trajectory in M_k using π_k.
  Update $\widehat{W}_{sa}$ and $\hat{\theta}_{sa}$ for observed samples.

The algorithm, thus, generates exploration policies by using perturbed rewards for planning. Similarly to Russo (2019), we can show the following bound for the expected regret incurred by GLM-RLSVI:

Theorem 4.7. *For any contextual MDP with given link function Φ, in Algorithm 3, if the MDP parameters for M_k are estimated using Algorithm 1, with reward bonuses $b_{k,h}(s,a) \sim N(0, SH\varphi_{k,h}(s,a))$ where $\varphi_{k,h}(s,a)$ is defined in Line. 11, the algorithm satisfies:*

$$\bar{R}(K) = \mathbb{E}\left[\sum_{k=1}^{K} V_k^* - V_k^{\pi_k}\right] = \widetilde{\mathcal{O}}\left(\left(\frac{\sqrt{d}\max_{s,a}\|W^*_{sa}\|_F}{\sqrt{\alpha}} + \frac{d}{\alpha}\right)\beta\sqrt{H^7 S^3 AK}\right)$$

The proof of the regret bound is given in Appendix B.2. Our regret bound is again worse by a factor of $\sqrt{H}$ when compared to the $\widetilde{\mathcal{O}}(H^3 S^{3/2}\sqrt{AK})$ bound from Russo (2019) for the tabular case. Therefore, such randomized bonus based exploration algorithms can also be used in the CMDP framework with similar regret guarantees as the tabular case.

5 LOWER BOUND FOR GLM CMDP

In this section, we show a regret lower bound by constructing a family of hard instances for the GLM-CMDP problem. We build upon the construction of Osband and Van Roy (2016) and Jaksch et al. (2010) for the analysis[7]:

Theorem 5.1. *For any algorithm* **A**, *there exists CMDP's with S states, A actions, horizon H and $K \geq dSA$ for logit and linear combination case, such that the expected regret of* **A** *(for any sequence of initial states $\in \mathcal{S}^K$) after K episodes is:*

$$\mathbb{E}[R(K; \mathbf{A}, M_{1:K}, s_{1:K})] = \Omega(H\sqrt{dSAK})$$

The lower bound has the usual dependence on MDP parameters in the tabular MDP case, with an additional $\mathcal{O}(\sqrt{d})$ dependence on the context dimension. Thus, our upper bounds have a gap of $\mathcal{O}(H\sqrt{dS})$ with the lower bound even in the arguably simpler case of Example 2.4.

6 IMPROVED CONFIDENCE SETS FOR STRUCTURED SPACES

In Section 3, we derived confidence sets for W^* for the case when it lies in a bounded set. However, in many cases, we have additional prior knowledge about the problem in terms of possible constraints over the set $\mathcal{W}$. For example, consider a healthcare scenario where the context vector contains the genomic encoding of the patient. For treating any ailment, it is fair to assume that the patient's response to the treatment and the progression in general depends on a few genes rather than the entire genome which suggests a sparse dependence of the transition model on the context vector x. In terms of the parameter W^*, this translates as complete columns of the matrix being zeroed out for the irrelevant indices. Thus, it is desirable to construct confidence sets which take this specific structure into account and give more problem dependent bounds.

In this section, we show that it is possible to convert a generic regret guarantee of an online learner to a confidence set. If the online learner adapts to the structure of $\mathcal{W}$, we would get the aforementioned improvement. The conversion proof presented here is reminiscent of the techniques used in Abbasi-Yadkori et al. (2012) and Jun et al. (2017) with close resemblance to the latter. For this section, we use X_t to denote the $t \times d$ shaped matrix with each row as x_i and C_t as $t \times S$ shaped matrix with each row i being $(W_i x_i)^{\top}$[8]. Also, set $\overline{W}_t := Z_{t+1}^{-1} X_t^{\top} C_t$. Using a similar notation as before, we can give the following guarantee.

Theorem 6.1 (Multinomial GLM Online-to-confidence set conversion)**.** *Assume that loss function l_i defined in eq.* (5) *is α-strongly convex with respect to Wx. If an online learning oracle takes in the sequence $\{x_i, y_i\}_{i=1}^t$, and produces outputs $\{W_i\}_{i=1}^t$ for an input sequence*

[7] The proof is deferred to the appendix due to space constraints.

[8] We again solely consider the estimation problem for a single (s,a) pair and study a t-indexed online estimation problem.

$\{x_i, y_i\}_{i=1}^t$, *such that:*

$$\sum_{i=1}^{t} l_i(W_i) - l_i(W) \leq B_t \quad \forall W \in \mathcal{W}, t > 0,$$

then with $\overline{W}_t$ *as defined above, with probability at least* $1-\delta$*, for all* $t \geq 1$*, we have*

$$\|W^* - \overline{W}_t\|^2_{Z_{t+1}} \leq \gamma_t$$

where $\gamma_t := \gamma'_t(B_t) + \lambda B^2 S - (\|C_t\|_F^2 - \langle \overline{W}_t, X_t^\top C_t \rangle)$,

$$\gamma'_t(B_t) := 1 + \tfrac{4}{\alpha} B_t + \tfrac{8}{\alpha^2} \log\left(\tfrac{1}{\delta}\sqrt{4 + \tfrac{8B_t}{\alpha} + \tfrac{16}{\alpha^4 \delta^2}}\right).$$

The complete proof can be found in Appendix E. Note that, all quantities required in the expression γ_t can be incrementally computed. The required quantities are Z_t and Z_t^{-1} along with $X_t^\top C_t$ which are incrementally updated with $O(\text{poly}(S, d))$ computation. Also, we note that this confidence set is meaningful when B_t is poly-logarithmic in t which is possible for strongly convex losses as shown in Jun et al. (2017). The dependence on S and d is the same as the previous construction, but the dependence on the strong convexity parameter is worse.

Column sparsity of W^* Similar to sparse stochastic linear bandit, as discussed in Abbasi-Yadkori et al. (2012), one can use an online learning method with the group norm regularizer ($\|W\|_{2,1}$). Therefore, if an efficient online no-regret algorithm has an improved dependence on the sparsity coefficient p, we can get an $O(\sqrt{p \log d})$ size confidence set. This will improve the final regret bound to $\tilde{\mathcal{O}}(\sqrt{pdT})$ as observed in the linear bandit case. To our knowledge, even in the sparse adversarial linear regression setting, obtaining an efficient and sparsity aware regret bound is an open problem.

7 DISCUSSION

Here, we discuss the obtained regret guarantees for our methods along with the related work. Further, we outline the algorithmic/analysis components which are different from the tabular MDP case and lead to interesting open questions for future work.

7.1 RELATED WORK

Contextual MDP To our knowledge, Hallak et al. (2015) first used the term contextual MDPs and studied the case when the context space is finite and the context is not observed during interaction. They propose `CECE`, a clustering based learning method and analyze its regret. Modi et al. (2018) generalized the CMDP framework and proved the PAC exploration bounds under smoothness and linearity assumptions over the contextual mapping. Their PAC bound is incomparable to our regret bound as a no-regret algorithm can make arbitrarily many mistakes $\Delta_k \geq \epsilon$ as long as it does so sufficiently less frequently.

Our work can be best compared with Abbasi-Yadkori and Neu (2014) and Dann et al. (2019) who propose regret minimizing methods for CMDPs. Abbasi-Yadkori and Neu (2014) consider an online learning scenario where the values $p_k(s'|s,a)$ are parameterized by a GLM. The authors give a no-regret algorithm which uses confidence sets based on Abbasi-Yadkori et al. (2012). However, their next state distributions are not normalized which leads to invalid next state distributions. Due to these modelling errors, their results cannot be directly compared with our analysis. Even if we ignore their modelling error, in the linear combination case, we get an $\tilde{\mathcal{O}}(S\sqrt{A})$ improvement. Similarly, Dann et al. (2019) proposed an OFU based method `ORLC-SI` for the linear combination case. Their regret bound is $\widetilde{\mathcal{O}}(\sqrt{S})$ worse than our bound for `GLM-ORL`. In addition, the work also showed that obtaining a finite mistake bound guarantees for such CMDPs requires a non-trivial and novel confidence set construction. In this paper, we show that a polylog(K) mistake bound can still be obtained. For a quick comparison, Table 1 shows the results from the two papers.

(Generalized) linear bandit Our reward model is based on the (stochastic) linear bandit problem first studied by Abe et al. (2003). Our work borrows key results from Abbasi-Yadkori et al. (2011) for both the reward estimator and during analysis for the GLM case. Extending the linear bandit problem, Filippi et al. (2010) first proposed the generalized linear contextual bandit setting and showed a $\mathcal{O}(d\sqrt{T})$ regret bound. We, however, leverage the approach from Zhang et al. (2016) and Jun et al. (2017) who also studied the logistic bandit and GLM Bernoulli bandit case. We extend their proposed algorithm and analysis to a generic categorical GLM setting. Consequently, our bounds also incur a dependence on the strong convexity parameter $\frac{1}{\alpha}$ of the GLM which was recently shown to be unavoidable by Foster et al. (2018) for proper learning in the closely related online logistic regression problem.

Regret analysis in tabular MDPs Auer and Ortner (2007) first proposed a no-regret online learning algorithm for average reward infinite horizon MDPs, and the problem has been extensively studied afterwards. More recently, there has been an increased focus on fixed horizon problems where the gap between the upper and lower bounds has been effectively closed. Azar et al. (2017) and Dann et al. (2019), both provide optimal regret guarantees

Algorithm	$R^{\text{Linear}}(K)$	$R^{\text{Logit}}(K)$	$P_x(\cdot\|s,a)$ **normalized**
Algorithm 1 (Abbasi-Yadkori and Neu, 2014)	$\widetilde{\mathcal{O}}(dH^3S^2A\sqrt{K})$	✗	✗
ORLC-SI (Dann et al., 2019)	$\widetilde{\mathcal{O}}(dH^2S^{3/2}\sqrt{AK})$	✗	✗
GLM-ORL (this work)	$\widetilde{\mathcal{O}}(dH^2S\sqrt{AK})$	$\widetilde{\mathcal{O}}(dH^2S^3\sqrt{AK})$	✓

Table 1: Comparison of regret guarantees for CMDPs. Last column denotes whether the transition dynamics $P_x(\cdot|s,a)$ are normalized in the model or not.

($\widetilde{\mathcal{O}}(H\sqrt{SAK})$) for tabular MDPs. Another series of papers (Osband et al., 2013, 2016; Russo et al., 2018) study Thompson sampling based randomized exploration methods and prove Bayesian regret bounds. Russo (2019) recently proved a worst case regret bound for RLSVI-style methods (Osband et al., 2016). The algorithm template and proof structure of GLM-RLSVI is borrowed from their work.

Feature-based linear MDP Yang and Wang (2019a) consider an RL setting where the MDP transition dynamics are low-rank. Specifically, given state-action features $\phi(s,a)$, they assume a setting where $p(s'|s,a) \coloneqq \sum_{i=1}^{d} \phi_i(s,a)\nu_i(s')$ where ν_i are d base distributions over the state space. This structural assumption guarantees that the $Q^\pi(s,a)$ value functions are linear in the state-action features for every policy. Yang and Wang (2019b); Jin et al. (2019) have recently proposed regret minimizing algorithms for the linear MDP setting. Although, their algorithmic structure is similar to ours (linear bandit based bonuses), the linear MDP setting is only superficially related to CMDP. In our case, the value functions are not linear in the contextual features for every policy and/or context. Thus, the two MDP frameworks and their regret analyses are incomparable.

7.2 CLOSING THE REGRET GAP

From the lower bound in Section 5, it is clear that the regret bound of GLM-ORL is sub-optimal by a factor of $\widetilde{\mathcal{O}}(H\sqrt{dS})$. As mentioned previously, for episodic MDPs, Azar et al. (2017) and Dann et al. (2019) propose minimax-optimal algorithms. The key technique in these analyzes is to directly build a confidence interval for the value functions and use a refined analysis using empirical Bernstein bonuses based on state-action visit counts saves a factor of $\mathcal{O}(\sqrt{HS})$. In our case, we use a Hoeffding style bonus for learning the next state distributions to derive confidence sets for the value function. Further, the value functions in GLM-CMDP do not have a nice structure as a function of the context variable and therefore, these techniques do not trivially extend to CMDPs. Similarly, the dependence on context dimension d is typically resolved by dividing the samples into phases which make them statistically independent (Auer, 2002; Chu et al., 2011; Li et al., 2017). However, for CMDPs, these filtering steps cannot be easily performed while ensuring long horizon optimistic planning. Thus, tightening the regret bounds for CMDPs is highly non-trivial and we leave this for future work.

8 CONCLUSION AND FUTURE WORK

In this paper, we have proposed optimistic and randomized no-regret algorithms for contextual MDPs which are parameterized by generalized linear models. We provide an efficient online Newton step (ONS) based update method for constructing confidence sets used in the algorithms. This work also outlines potential future directions: close the regret gap for tabular CMDPs, devise an efficient and sparsity aware regret bound and investigate whether a near-optimal mistake and regret bound can be obtained simultaneously. Lastly, extension of the framework to non-tabular MDPs is an interesting problem for future work.

Acknowledgements

AM thanks Satinder Singh and Alekh Agarwal for helpful discussions. This work was supported in part by a grant from the Open Philanthropy Project to the Center for Human-Compatible AI, and in part by NSF grant CAREER IIS-1452099. AT would like to acknowledge the support of a Sloan Research Fellowship.

References

Abbasi-Yadkori, Y. and Neu, G. (2014). Online learning in MDPs with side information. *arXiv preprint arXiv:1406.6812*.

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320.

Abbasi-Yadkori, Y., Pal, D., and Szepesvari, C. (2012).

Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Artificial Intelligence and Statistics*, pages 1–9.

Abe, N., Biermann, A. W., and Long, P. M. (2003). Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica*, 37(4):263–293.

Agarwal, A. (2013). Selective sampling algorithms for cost-sensitive multiclass prediction. In *International Conference on Machine Learning*, pages 1220–1228.

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422.

Auer, P. and Ortner, R. (2007). Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 49–56.

Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org.

Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214.

Dann, C., Li, L., Wei, W., and Brunskill, E. (2019). Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pages 1507–1516.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. (2010). Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594.

Foster, D. J., Kale, S., Luo, H., Mohri, M., and Sridharan, K. (2018). Logistic regression: The importance of being improper. In *Conference On Learning Theory*, pages 167–208.

Hallak, A., Di Castro, D., and Mannor, S. (2015). Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*.

Hazan, E., Agarwal, A., and Kale, S. (2007). Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192.

Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2019). Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388*.

Jun, K.-S., Bhargava, A., Nowak, R., and Willett, R. (2017). Scalable generalized linear bandits: Online computation and hashing. In *Advances in Neural Information Processing Systems*, pages 99–109.

Lattimore, T. and Szepesvri, C. (2020). *Bandit Algorithms*. Cambridge University Press.

Lauritzen, S. L. (1996). *Graphical models*, volume 17. Clarendon Press.

Li, L., Lu, Y., and Zhou, D. (2017). Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080.

Modi, A., Jiang, N., Singh, S., and Tewari, A. (2018). Markov decision processes with continuous side information. In *Algorithmic Learning Theory*, pages 597–618.

Osband, I., Russo, D., and Van Roy, B. (2013). (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011.

Osband, I. and Van Roy, B. (2016). On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*.

Osband, I., Van Roy, B., and Wen, Z. (2016). Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pages 2377–2386.

Russo, D. (2019). Worst-case regret bounds for exploration via randomized value functions. In *Advances in Neural Information Processing Systems*, pages 14410–14420.

Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., Wen, Z., et al. (2018). A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96.

Yang, L. and Wang, M. (2019a). Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004.

Yang, L. F. and Wang, M. (2019b). Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. *arXiv preprint arXiv:1905.10389*.

Zhang, L., Yang, T., Jin, R., Xiao, Y., and Zhou, Z.-h. (2016). Online stochastic linear optimization under one-bit feedback. In *International Conference on Machine Learning*, pages 392–401.

A PROOF OF THEOREM 3.1

We closely follow the analysis from Zhang et al. (2016) and use properties of the categorical output space to adapt it to our case. The analysis is fairly similar, but carefully manipulating the matrix norms saves a factor of $\mathcal{O}(S)$ in the confidence widths. For notation, we use $\nabla l_t(W_t)$ to refer to the derivative with respect to the matrix for loss l_t and $\nabla l_t(W_t x_t)$ for the derivative with respect to the projection $W_t x_t$. B_p denotes the upper bound on the ℓ_2-norm of each row $W^{(i)}$ and R is the assumed bound on the context norm $\|x\|_2$. Now, using the strong convexity of the loss function l_t with respect to $W_t x_t$, for all t, we have:

$$l_t(W_t) - l_t(W^*) \leq \langle \nabla l_t(W_t x_t), W_t x_t - W^* x_t \rangle - \tfrac{\alpha}{2} \underbrace{\|W^* x_t - W_t x_t\|_2^2}_{:=b_t}$$

Taking expectation with respect to the categorical sample y_t, we get:

$$\begin{aligned} 0 &\leq \mathbb{E}_{y_t}[l_t(W_t) - l_t(W^*)] \\ &\leq \mathbb{E}_{y_t}[\langle \nabla l_t(W_t x_t), W_t x_t - W^* x_t \rangle] - \tfrac{\alpha}{2} b_t \\ &\leq \mathbb{E}_{y_t}[\langle \nabla l_t(W_t x_t), W_t x_t - W^* x_t \rangle] - \tfrac{\alpha}{2} b_t \end{aligned} \tag{16}$$

where the `lhs` is obtained by using the calibration property from eq. (6). Now, for the first term on `rhs`, we have:

$$\begin{aligned} &\mathbb{E}_{y_t}[\langle \nabla l_t(W_t x_t), W_t x_t - W^* x_t \rangle] \\ &= \mathbb{E}_{y_t}[\langle \nabla \Phi(W_t x_t) - y_t, W_t x_t - W^* x_t \rangle] \\ &= (\tilde{p}_t - p_t)^\top (W_t - W^*) x_t \\ &= \underbrace{(\tilde{p}_t - y_t)^\top (W_t - W^*) x_t}_{:=\mathbf{I}} \\ &\quad + \underbrace{(y_t - p_t)^\top (W_t - W^*) x_t}_{:=c_t} \end{aligned} \tag{17}$$

where $\tilde{p}_t = \nabla \Phi(W_t x_t)$ and $\mathbb{E}[y_t] = p_t = \nabla \Phi(W^* x_t)$. We bound the term $\mathbf{I}$ using the following lemma:

Lemma A.1.

$$\begin{aligned} &\langle \nabla l_t(W_t x_t), W_t x_t - W^* x_t \rangle \\ &\leq \frac{\|W_t - W^*\|_{Z_{t+1}}}{2\eta} - \frac{\|W_{t+1} - W^*\|_{Z_{t+1}}}{2\eta} \\ &\quad + 2\eta \|x_t\|^2_{Z^{-1}_{t+1}} \end{aligned} \tag{18}$$

Proof. To prove this, we go back to the update rule in (7) which has the following form:

$$Y = \arg\min_{W \in \mathcal{W}} \frac{\|W - X\|_M^2}{2} + \eta a^\top W b$$

with $Y = W_{t+1}$, $X = W_t$, $a = \nabla l_t(W_t x_t) = \tilde{p}_t - y_t$, $b = x_t$ and $M = Z_{t+1}$. For a solution to any such optimization problem, by the first order optimality conditions, we have:

$$\begin{aligned} \langle (Y - X)M + \eta a b^\top, W - Y \rangle &\geq 0 \\ (Y - X)MW &\geq (Y - X)MY \\ &\quad - \eta a^\top (W - Y) b \end{aligned}$$

Using this first order condition, we have

$$\begin{aligned} &\|X - W\|_M^2 - \|Y - W\|_M^2 \\ &= \sum_{i=1}^{S} X^i M X^i + W^i M W^i - Y^i M Y^i \\ &\quad - W^i M W^i + 2(Y^i - X^i) M W^i \\ &\geq \|X - Y\|_M^2 - 2\eta a^\top (W - Y) b \\ &= \|X - Y\|_M^2 + 2\eta a^\top (Y - X) b \\ &\quad - 2\eta a^\top (W - X) b \\ &\geq \arg\min_{A \in \mathbb{R}^{S \times d}} \|A\|_M^2 + 2\eta a^\top A b - 2\eta a^\top (W - X) b \end{aligned} \tag{19}$$

Noting that $a = \tilde{p}_t - y^t$, we get

$$\begin{aligned} \arg\min_{A \in \mathbb{R}^{S \times d}} \|A\|_M^2 + 2\eta a^\top A b &\geq \sum_{i=1}^{S} -\eta^2 a_i^2 \|b\|^2_{M^{-1}} \\ &\geq -4\eta^2 \|b\|^2_{M^{-1}} \end{aligned}$$

Substituting this and $W = W^*$ along with other terms in ineq. (19) proves the stated lemma (ineq. (18)). □

Thus, from eqs. (16), (17) and (18), we have

$$\begin{aligned} &\|W_{t+1} - W^*\|_{Z_{t+1}} \\ &\leq \|W_t - W^*\|_{Z_t} - \frac{\eta\alpha}{2} b_t + 2\eta c_t + 4\eta^2 \|x_t\|^2_{Z^{-1}_{t+1}} \end{aligned} \tag{20}$$

Bounding the first term on the `rhs` similarly, and telescoping the sum, we get:

$$\begin{aligned} &\|W_{t+1} - W^*\|_{Z_{t+1}} + \frac{\eta\alpha}{2} \sum_{i=1}^{t} b_i \\ &\leq \|W^*\|_{Z_1} + 2\eta \sum_{i=1}^{t} c_i + 4\eta^2 \sum_{i=1}^{t} \|x_i\|^2_{Z^{-1}_{i+1}} \\ &\leq \lambda \|W^*\|_F^2 + 2\eta \sum_{i=1}^{t} c_i + 4\eta^2 \sum_{i=1}^{t} \|x_i\|^2_{Z^{-1}_{i+1}} \end{aligned} \tag{21}$$

We will now bound the sum $\sum_{i=1}^{t} c_i$ in ineq. (21) using Bernstein's inequality for martingales in the same manner as Zhang et al. (2016):

Lemma A.2. *With probability at least* $1-\delta$*, we have:*

$$\sum_{i=1}^{t} c_i \leq 4B_pR + \frac{\alpha}{4}\sum_{i=1}^{t} b_i + \left(\frac{4}{\alpha} + \frac{8B_pR}{3}\right)\tau_t \quad (22)$$

where $\tau_t = \log(2\lceil 2\log St\rceil t^2/\delta)$.

Proof. The result can be easily derived from the proof of Lemma 5 in Zhang et al. (2016). We provide the key steps here for completeness.

We first note that c_t is a martingale difference sequence with respect to filtration $\mathcal{F}_t$ induced by the first t rounds including the next context x_{t+1}:

$$\begin{aligned}&\mathbb{E}\left[(y_t - p_t)^\top (W_t - W^*)x_t | \mathcal{F}_{t-1}\right] \\ &= \mathbb{E}\left[(y_t - p_t)|\mathcal{F}_{t-1}\right]^\top (W_t - W^*)x_t = 0\end{aligned}$$

Further, each term in this martingale series can be bounded as:

$$\begin{aligned}|c_t| &= (y_t - p_t)^\top (W_t - W^*)x_t \\ &\leq \|(y_t - p_t)\|_1 \|(W_t - W^*)x_t\|_\infty \\ &\leq 4B_pR\end{aligned}$$

Similarly, for martingale $C_t := \sum_{i=1}^{t} c_i$, we bound the conditional variance as

$$\begin{aligned}\Sigma_t^2 &= \sum_{i=1}^{t} \mathbb{E}_{y_i}\left[\left((y_t - p_t)^\top (W_t - W^*)x_t\right)^2\right] \\ &\leq \sum_{i=1}^{t} \mathbb{E}_{y_i}\left[\left(y_t^\top (W_t - W^*)x_t\right)^2\right] \\ &\leq \underbrace{\sum_{i=1}^{t} \|(W_t - W^*)x_t\|_2^2}_{:=A_t}\end{aligned}$$

Thus, we have a natural upper bound for the conditional variance which is $\Sigma_t^2 \leq 4B_p^2R^2St$. Now, consider two scanarios: `CASE I`: $A_t \geq 4B_p^2R^2/St$ and `CASE II`: $4B_p^2R^2/St \leq A_t \leq 4B_p^2R^2St$.

`CASE I`: Here, we directly bound the sum as

$$\begin{aligned}C_t \leq \sum_{i=1}^{t} |c_i| &\leq 2\sum_{i=1}^{t} \|(W_t - W^*)x_t\|_2 \\ &\leq 2\sqrt{t\sum_{i=1}^{t} \|(W_t - W^*)x_t\|_2^2} \leq 4B_pR\end{aligned}$$

`CASE II`: We directly use the expression after applying Bernstein's inequality along with the peeling technique from Zhang et al. (2016). Using that, we have:

$$\begin{aligned}&P\left[C_t \geq 2\sqrt{A_t\tau_t} + \frac{8B_pR\tau_t}{3}\right] \\ &\leq \sum_{j=-\log S}^{m} P\Big[C_t \geq 2\sqrt{A_t\tau_t} + \frac{8B_pR\tau_t}{3}, \\ &\qquad \frac{4B_pR^2 2^j}{t} \leq A_t \leq \frac{4B_pR^2 2^{j+1}}{t}\Big] \\ &\leq m' e^{-\tau_t}\end{aligned}$$

where $m = \log St^2$ and $m' = m + \log S = \log S^2t^2$. We set $\tau_t = \log\frac{2m't^2}{\delta}$, we get that with probability at least $1 - \delta/2t^2$, we have:

$$C_t \leq 2\sqrt{A_t\tau_t} + \frac{8B_pR\tau_t}{3}$$

Taking a union bound over $t \geq 0$ and substituting $A_t = \sum_{i=1}^{t} b_i$, with probability at least $1-\delta$, for all $t \geq 0$, we get:

$$\sum_{i=1}^{t} c_i \leq 4B_pR + 2\sqrt{\tau_t \sum_{i=1}^{t} b_i} + \tfrac{8B_pR}{3}\tau_t$$

Using the RMS-AM inequality, we get the desired expression:

$$\sum_{i=1}^{t} c_i \leq 4B_pR + \frac{\alpha}{4}\sum_{i=1}^{t} b_i + \left(\frac{4}{\alpha} + \frac{8B_pR}{3}\right)\tau_t$$

□

Substituting the high probability upper bound over $\sum_{i=1}^{t} c_i$ in eq. (21), we get:

$$\begin{aligned}&\|W_{t+1} - W^*\|_{Z_{t+1}} \\ &\leq \lambda\|W^*\|_F^2 + 2\eta\Big[4B_pR + \Big(\frac{4}{\alpha} + \frac{8}{3}B_pR\Big)\tau_t\Big] \\ &\quad + 4\eta^2\sum_{i=1}^{t} \|x_t\|_{Z_{t+1}^{-1}}^2 \qquad (23)\end{aligned}$$

For getting the final result, we now bound the elliptic potential using the following Lemma from Zhang et al. (2016):

Lemma A.3 (Lemma 6, Zhang et al. (2016))**.**

$$\sum_{i=1}^{t} \|x_t\|_{Z_{t+1}^{-1}}^2 \leq \frac{2}{\eta\alpha}\log\frac{\det(Z_{t+1})}{\det(Z_1)}$$

B REGRET ANALYSIS

B.1 PROOF OF THEOREM 4.1

We now provide a complete proof of Theorem 4.1.

B.1.1 Failure events and bounding failure probabilities

To begin with, we write the important failure events for the algorithm $F = F^{(r)} \cup F^{(p)} \cup F^{(O)}$ where each sub-event is defined as follows:

$$F^{(O)} := \Big\{\exists K \in \mathbb{N} : \sum_{k=1}^{K} \sum_{h,s,a} \Big(\mathbb{P}_k[s_h, a_h = s, a | s_{k,1}] - \mathbb{I}[s_{k,h} = s, a_{k,h} = a]\Big) \geq SH\sqrt{K \log \tfrac{6\log(2K)}{\delta_1}}\Big\}$$

$$F^{(p)} := \Big\{\exists\, s \in \mathcal{S}, a \in \mathcal{A}, k \in \mathbb{N} : \|W_{sa} - \widehat{W}_{k,sa}\|_{Z_{k,sa}} \geq \sqrt{\gamma_{k,sa}}\Big\}$$

$$F^{(r)} := \Big\{\exists\, s \in \mathcal{S}, a \in \mathcal{A}, k \in \mathbb{N} : \|\theta_{sa} - \widehat{\theta}_{k,sa}\|_{Z_{k,sa}} \geq \zeta_{k,sa}\Big\}$$

Using high-probability guarantees for parameter estimation and concentration of measure, we have the guarantee that:

Lemma B.1. *The probabilities for failure events $F^{(O)}, F^{(p)}$ and $F^{(r)}$ are bounded bounded by $SH\delta_1$, $SA\delta_p$ and $SA\delta_r$ respectively.*

Proof. The guarantee for $F^{(p)}$ follows from Theorem 3.1 in Section 3. The failure probability $P(F^{(r)})$ can be bounded by using Theorem 20.5 from Lattimore and Szepesvri (2020).

Lastly, the failure probability $P(F^{(O)})$ is directly taken from Lemma 23 of Dann et al. (2019). □

B.1.2 Regret incurred outside failure events

Lemma B.2 (Optimism)**.** *If all the confidence intervals as computed in Algorithm 2 are valid for all episodes k, then outside of failure event F, for all k and $h \in [H]$ and $s, a \in \mathcal{S} \times \mathcal{A}$, we have:*

$$\tilde{Q}_{k,h}(s,a) \geq Q^*_{k,h}(s,a)$$

Proof. For every episode, the lemma is true trivially for $H+1$. Assume that it is true for $h+1$. For h, we have:

$$\begin{aligned}
&\tilde{Q}_{k,h}(s,a) - Q^*_{k,h}(s,a) \\
&= (\widehat{P}_k(s,a)^\top \tilde{V}_{k,h+1} + \hat{r}_k(s,a) + \varphi_{k,h}(s,a)) \wedge V_h^{\max} \\
&\quad - P_k(s,a)^\top V^*_{k,h+1} - r_k(s,a) \\
&= \hat{r}_k(s,a) - r_k(s,a) + \widehat{P}_k(s,a)^\top (\tilde{V}_{k,h+1} - V^*_{k,h+1}) \\
&\quad + \varphi_{k,h}(s,a) - (P_k(s,a) - \widehat{P}_k(s,a))^\top V^*_{k,h+1} \\
&\geq -|\hat{r}_k(s,a) - r_k(s,a)| + \varphi_{k,h}(s,a) \\
&\quad - \|P_k(s,a) - \widehat{P}_k(s,a)\|_1 \|\tilde{V}_{k,h+1}\|_\infty \geq 0
\end{aligned}$$

In the second equality step, we use the fact that when $\tilde{Q}_{k,h}(s,a) = V_h^{\max}$, the requirement is trivially satisfied. When $\tilde{Q}_{k,h}(s,a) < V_h^{\max}$, the step follows by definition. The last step uses the guarantee on confidence intervals and the inductive assumption for $h+1$. Therefore, the estimated Q-values are optimistic by induction. □

Therefore, using the optimism guarantee, we can bound the instantaneous regret Δ_k in episode k as: $V^*_{k,1}(s) - V^{\pi_k}_{k,1}(s) \leq \widetilde{V}_{k,1}(s) - V^{\pi_k}_{k,1}(s)$. Thus, we have:

$$\begin{aligned}
\Delta_k &\leq \widetilde{V}_{k,1}(s) - V^{\pi_k}_{k,1}(s) \\
&\leq (\widehat{P}_k(s,a)^\top \widetilde{V}_{k,2} + \hat{r}_k(s,a) + \varphi) \wedge V_1^{\max} \\
&\quad - P_k(s,a)^\top V^{\pi_k}_{k,2} - r_k(s,a) \\
&\leq (\varphi + \widehat{P}_k(s,a) - P_k(s,a))^\top \widetilde{V}_{k,2} + \hat{r}_k(s,a) \\
&\quad - r_k(s,a)) \wedge V_1^{\max} + P_k(s,a)^\top (V^{\pi_k}_{k,2} - \widetilde{V}_{k,2}) \\
&\leq 2\varphi \wedge V_1^{\max} + P_k(s,a)^\top (V^{\pi_k}_{k,2} - \widetilde{V}_{k,2}) \\
&\leq \sum_{h,s,a} \Big[\mathbb{P}_k[s_h, a_h = s, a | s_{k,1}] \\
&\qquad (2\varphi(s,a) \wedge V_h^{\max})\Big] \qquad (24)
\end{aligned}$$

Using Lemma B.1.1, we can show the following result:

Lemma B.3. *Outside the failure event $F^{(O)}$, i.e., with probability at least $1 - SH\delta_1$, the total regret $R(K)$ can be bounded by*

$$\begin{aligned}
R(K) &\leq SH^2\sqrt{K \log \tfrac{6\log 2K}{\delta_1}} \\
&\quad + \sum_{k=1}^{K} \sum_{h=1}^{H} \cdot (2\varphi_{k,h}(s_{k,h}, a_{k,h}) \wedge V_h^{\max}) \qquad (25)
\end{aligned}$$

Proof.

$$\begin{aligned}\Delta_k &\leq \sum_{h,s,a} [\mathbb{P}_k[s_h, a_h = s, a|s_{k,1}](2\varphi(s,a) \wedge V_h^{\max})] \\ &\leq \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s,a}\Big(\mathbb{P}_k[s_h, a_h = s, a|s_{k,1}] \\ &\quad - \mathbb{I}_{k,h}(s,a)\Big)(2\varphi(s_{k,h}, a_{k,h}) \wedge V_h^{\max}) \\ &\quad + \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{I}_{k,h}(s,a)(2\varphi(s_{k,h}, a_{k,h}) \wedge V_h^{\max})\end{aligned}$$

where $\mathbb{I}_{k,h}(s,a)$ is the indicator function $\mathbb{I}[s_{k,h} = s, a_{k,h} = a]$. From Lemma B.1.1, we know that the first term is bounded by $SH\sqrt{K \log \frac{6 \log 2K}{\delta_1}}$ with probability at least $1 - SH\delta_1$. □

Before bounding the second term in ineq. (25), we state the following Lemma from Abbasi-Yadkori et al. (2011) which is used frequently in our analysis:

Lemma B.4 (Determinant-Trace inequality)**.** *Suppose $X_1, X_2, \ldots, X_t \in \mathbb{R}^d$ and for any $1 \leq s \leq t$, $\|X_s\|_2 \leq L$. Let $V_t \coloneqq \lambda\mathbf{I} + \sum_{s=1}^{t} X_s X_s^\top$ for some $\lambda \geq 0$. Then, we have:*

$$\det(V_t) \leq \left(\lambda + tL^2/d\right)^d$$

The second term in ineq. (12) can now be bounded as follows:

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{h=1}^{H}(2\varphi(s_{k,h}, a_{k,h}) \wedge V_h^{\max}) \\ &\leq \sum_{k=1}^{K}\sum_{h=1}^{H}(2\xi^{(r)}_{k,s_{k,h},a_{k,h}} \wedge V_h^{\max}) \\ &\quad + \sum_{k=1}^{K}\sum_{h=1}^{H}(2V_{h+1}^{\max}\xi^{(p)}_{k,s_{k,h},a_{k,h}} \wedge V_h^{\max}) \qquad (26)\end{aligned}$$

We ignore the reward estimation error in eq. (26) as it leads to lower order terms. The second expression can be again bounded as follows:

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{h=1}^{H}(2V_{h+1}^{\max}\xi^{(p)}_{k,s_{k,h},a_{k,h}} \wedge V_h^{\max}) \\ &\leq 2\sum_{k,h} V_h^{\max}\left(1 \wedge \beta\sqrt{S\gamma_k(s_{k,h},a_{k,h})}\|x_k\|_{Z^{-1}_{k,sa,h}}\right) \qquad (27)\end{aligned}$$

Using Lemma B.4, we see that

$$\begin{aligned}\gamma_k(s,a) &\coloneqq f_\Phi(k,\delta_p) + \frac{8\eta}{\alpha}\log\frac{\det(Z_{k,sa})}{\det(Z_{1,sa})} \\ &\leq \frac{\eta\alpha}{2S} + f_\Phi(KH,\delta_p) + \frac{8\eta}{\alpha}\log\frac{\det(Z_{K+1,sa})}{\det(Z_{1,sa})} \\ &\leq \frac{\eta\alpha}{2S} + f_\Phi(KH,\delta_p) + \frac{8\eta d}{\alpha}\log\left(1 + \frac{KHR^2}{\lambda d}\right)\end{aligned}$$

We use $f_\Phi(k,\delta_p)$ to refer to the Z_k independent terms in eq. (9). Setting $\bar{\gamma}_K$ to the last expression guarantees that $\frac{2S\bar{\gamma}_K}{\eta\alpha} \geq 1$. We can now bound the term in eq. (27) as:

$$\begin{aligned}&2\beta V_1^{\max}\sqrt{\frac{2S\bar{\gamma}_K}{\eta\alpha}}\sum_{k,h}\left(1 \wedge \sqrt{\frac{\eta\alpha}{2}}\|x_k\|_{Z^{-1}_{k,sa,h}}\right) \\ &\leq 2\beta V_1^{\max}\sqrt{\frac{2S\bar{\gamma}_K KH}{\eta\alpha}}\sqrt{\sum_{k,h}\left(1 \wedge \frac{\eta\alpha}{2}\|x_k\|^2_{Z^{-1}_{k,sa,h}}\right)} \qquad (28)\end{aligned}$$

Ineq. (28) follows by using Cauchy-Schwarz inequality. We now bound the elliptic potential inside the square root in ineq. (28):

Lemma B.5. *For any $K \in \mathbb{N}$, we have:*

$$\sum_{k,h}\left(1 \wedge \frac{\eta\alpha}{2}\|x_k\|^2_{Z^{-1}_{k,sa,h}}\right) \leq 2H\sum_{s,a}\log\left(\frac{\det Z_{k+1,sa}}{\det Z_{k,sa}}\right)$$

Proof. Note that, instead of summing up the weighted operator norm with changing values of $Z_{k,h}$ for each observed transition of a pair (s,a), we keep the matrix same for all observations in an episode. Note that, Z_k denotes the matrix at the beginning of episode k and therefore, does not include the terms $x_k x_k^\top$. Thus, for any episode k:

$$\begin{aligned}&\sum_{h=1}^{H}\left(1 \wedge \frac{\eta\alpha}{2}\|x_k\|^2_{Z^{-1}_{k,sa,h}}\right) \\ &\leq 2\sum_{s,a}\sum_{h=1}^{H}\mathbb{I}_{k,h}(s,a)\log\left(1 + \frac{\eta\alpha}{2}\|x_k\|^2_{Z^{-1}_{k,sa}}\right) \\ &= 2\sum_{s,a}N_k(s,a)\log\left(1 + \frac{\eta\alpha}{2}\|x_k\|^2_{Z^{-1}_{k,sa}}\right) \\ &\leq 2\sum_{s,a}N_k(s,a)\log\left(1 + N_k(s,a)\frac{\eta\alpha}{2}\|x_k\|^2_{Z^{-1}_{k,sa}}\right) \\ &= 2H\sum_{s,a}\log\left(\frac{\det Z_{k+1,sa}}{\det Z_{k,sa}}\right)\end{aligned}$$

where in the last step, we have used the following:

$$Z_{k+1} = Z_k^{1/2}\left(1 + \frac{\eta\alpha}{2}N_k Z_k^{-1/2}x_k x_k^\top Z_k^{-1/2}\right)Z_k^{1/2}$$

and then bound the determinant ratio using

$$\det Z_{k+1} = \det Z_k \left(1 + N_k \frac{\eta\alpha}{2} \|x_k\|^2_{Z_k^{-1}}\right)$$

□

Finally, by using Lemma B.4, we can bound the term as

$$\sum_{k=1}^{K}\sum_{h=1}^{H}(2V_{h+1}^{\max}\xi^{(p)}_{k,s_{k,h},a_{k,h}} \wedge V_h^{\max})$$
$$\leq 4\beta V_1^{\max}\sqrt{\frac{2S\bar{\gamma}_K KH}{\eta\alpha}}\sqrt{2HSAd\log\left(1+\frac{KHR^2}{\lambda d}\right)}$$

Now, we set each individual failure probability$\delta_1 = \delta_p = \delta_r = \delta/(2SA+SH)$. Upon taking a union bound over all events, we get the total failure probability as δ. Therefore, with probability at least $1-\delta$, we can bound the regret of `GLM-ORL` as

$$R(K) = \widetilde{\mathcal{O}}\left(\left(\frac{\sqrt{d}\max_{s,a}\|W^*_{sa}\|_F}{\sqrt{\alpha}} + \frac{d}{\alpha}\right)\beta SH^2\sqrt{AK}\right)$$

where $\max_{s,a}\|W^*_{sa}\|_F$ is replaced by the problem dependent upper bound assumed to be known apriori.

B.2 PROOF OF THEOREM 4.7

Our analysis will closely follow the proof from Russo (2019). We start by writing the concentration result for estimating MDP M_k by using Algorithm 1 and the linear bandit estimators. For notation, we use $\widehat{M}_k$ to denote the MDP constructed using the estimates $\widehat{W}_k$ and $\widehat{\theta}_k$. The perturbed MDP used in the algorithm is denoted by $\overline{M}_k$ and $\widetilde{M}_k$ will denote an MDP constructed using another set of *i.i.d.* reward bonuses as $\overline{M}_k$. Specifically, we have:

Lemma B.6. *Let $\mathcal{M}_k$ be the following set of MDPs:*

$$\mathcal{M}_k \coloneqq \{(P', R') : \forall(h,s,a),\ |(R'(s,a) - R_k(s,a)) + \langle P'(s,a) - P_k(s,a), V_{k,h+1}\rangle| \leq \varphi_{k,h}(s,a)\}$$

where $\varphi^2_{k,h}(s,a) = (\beta\sqrt{S\gamma_{k,sa}}(H-h) + \zeta_{k,sa})\|x_k\|_{Z^{-1}_{k,sa}}$. If we choose $\delta_p = \delta_r = \pi^2/SA$, then, we have:

$$\sum_{k\in\mathbb{N}} P_k[\widehat{M}_k \notin \mathcal{M}_k] \leq \frac{\pi^2}{6}$$

Proof. The proof follows from the analysis in Appendix B.1 where the union bound over all (s,a) pairs gives the total failure probability to be $\frac{\pi^2}{6}$. □

Given the concentration result, Lemma 4 from Russo (2019) directly applies to the CMDP setting in the following form:

Lemma B.7. *Let π^*_k be the optimal policy for MDP M_k. If $\widehat{M}_k \in \mathcal{M}_k$ and reward bonuses $b_{k,h}(s,a) \sim N(0, HS\varphi^2_{k,h}(s,a))$, then we have*

$$P\left[v^{\pi_k}_{\overline{M}_k} \geq v^{\pi^*}_{M_k} | \mathcal{H}_{k-1}\right] \geq \mathbb{F}(-1)$$

where $\widehat{M}_k$ is the estimated MDP, $\overline{M}_k$ is the MDP obtained after perturbing the rewards and $\mathbb{F}(\cdot)$ is the cdf for the standard normal distribution.

In a similar fashion, the following result can also be easily verified:

Lemma B.8. *For an absolute constant $c = \mathbb{F}(-1)^{-1} \leq 6.31$, we have:*

$$R(K) \coloneqq \mathbb{E}_{Alg}\left[\sum_{k=1}^{K} v^*_k(s_{k,1}) - v^{\pi_k}_k(s_{k,1})\right]$$
$$\leq (c+1)\mathbb{E}\left[\sum_{k=1}^{K}\left|v^{\pi_k}_{\overline{M}_k} - v^{\pi_k}_{M_k}\right|\right]$$
$$+ c\mathbb{E}\left[\sum_{k=1}^{K}\left|v^{\pi_k}_{\widetilde{M}_k} - v^{\pi_k}_{M_k}\right|\right] + H\frac{\pi^2}{6}$$

We will now bound the first term on the `rhs` of Lemma B.8 to get the final regret bound. The second term can be bounded in the same manner. For each episode, the summand in the first term can be written as:

$$v^{\pi_k}_{\overline{M}}(s_{k,1}) - v^{\pi_k}_{M_k}(s_{k,1})$$
$$= \Big|\mathbb{E}\Big[\sum_{h=1}^{H}\Big(\langle P_k(s_{k,h},a_{k,h}) - \widehat{P}_k(s_{k,h},a_{k,h}), \overline{V}_{k,h+1}\rangle + \hat{r}_k(s_{k,h},a_{k,h}) - r_k(s_{k,h},a_{k,h}) + b_{k,h}(s_{k,h},a_{k,h})\Big)\Big|\mathcal{H}_{k-1}\Big]\Big|$$
$$\leq \left|\mathbb{E}\left[\sum_{h=1}^{H}\left\langle P_k(s_{k,h},a_{k,h}) - \widehat{P}_k(s_{k,h},a_{k,h}), \overline{V}_{k,h+1}\right\rangle\right] + \mathbb{E}\left[\sum_{h=1}^{H} r_k(s_{k,h},a_{k,h}) - \hat{r}_k(s_{k,h},a_{k,h})|\mathcal{H}_{k-1}\right]\right|$$
$$+ \mathbb{E}\left[\sum_{h=1}^{H}|b_{k,h}(s_{k,h},a_{k,h})|\,\Big|\mathcal{H}_{k-1}\right] \tag{29}$$

where $\overline{V}_{k,h+1}$ denotes the h^{th}-step value of policy π_k in $\overline{M}_k$. We will now bound each term individually where we ignore the reward term and the variance component

due to reward uncertainty as both lead to lower order terms. Specifically, we directly consider $\varphi^2_{k,h}(s,a) = 2\left(\beta\sqrt{S\gamma_{k,sa}}(H-h)\right)\|x_k\|_{Z^{-1}_{k,sa}}$. For the last expression in eq. (29), we focus on the first and third terms (the reward bonuses lead to lower order terms in the final regret bound).

Lemma B.9. *We have:*

$$\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}|b_{k,h}(s_{k,h},a_{k,h})|\,\middle|\,\mathcal{H}_{k-1}\right] = \widetilde{\mathcal{O}}\left(\left(\frac{\sqrt{d}\max_{s,a}\|W^*_{sa}\|_F}{\sqrt{\alpha}}+\frac{d}{\alpha}\right)\beta S^{3/2}H^{5/2}\sqrt{AK}\right)$$

Proof. We write $b_{k,h}(s_{k,h},a_{k,h}) = \sqrt{HS}\varphi_{k,h}(s_{k,h},a_{k,h})\xi_{k,h}(s_{k,h},a_{k,h})$ where $\xi_{k,h}(s_{k,h},a_{k,h}) \sim N(0,1)$. Therefore, by using Holder's inequality, we have:

$$\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}|b_{k,h}(s_{k,h},a_{k,h})|\,\middle|\,\mathcal{H}_{k-1}\right] \leq \mathbb{E}\left[\max_{k,h,s,a}\xi_{k,h}(s,a)\right]\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{HS}\varphi_{k,h}(s_{k,h},a_{k,h})\right]$$

By using (sub)-Gaussian maximal inequality, we know that

$$\mathbb{E}\left[\max_{k,h,s,a}\xi_{k,h}(s,a)\right] = \mathcal{O}\left(\log(HSAK)\right) \tag{30}$$

For the second expression, we have:

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{HS}\varphi_{k,h}(s_{k,h},a_{k,h})\right]\\ &\leq\sqrt{HS}\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\varphi_{k,h}(s_{k,h},a_{k,h})\right]\\ &\leq 2H^{3/2}\sqrt{S}\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}1\wedge\left(\beta\sqrt{S}\sqrt{\gamma_{k,sa}}\|x_k\|_{Z^{-1}_{k,sa}}\right)\right]\end{aligned}$$

where we used the definition of $\overline{\xi}^{(p)}_{k,h}$ used in Section 4.2. Using the upper bound above along with Lemmas B.4 and B.5, we obtain the bound:

$$\mathbb{E}\left[\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{HS}\varphi_{k,h}(s_{k,h},a_{k,h})\right] = \mathcal{O}\left(\beta H^{5/2}S^{3/2}\sqrt{\frac{dA\bar{\gamma}_K K}{\eta\alpha}}\sqrt{\log\left(1+\frac{KHR^2}{\lambda d}\right)}\right) \tag{31}$$

We get the final bound on the term by combining eqs. (30) and (31). □

We now bound the first term in eq. (29):

Lemma B.10. *With the ONS estimation method and the used randomized bonus, we have:*

$$\mathbb{E}\left[\sum_{k,h}\left|\langle P_k(s_{k,h},a_{k,h})-\widehat{P}_k(s_{k,h},a_{k,h}),\overline{V}_{k,h+1}\rangle\right|\right] = \widetilde{\mathcal{O}}\left(\left(\frac{\sqrt{d}\max_{s,a}\|W^*_{sa}\|_F}{\sqrt{\alpha}}+\frac{d}{\alpha}\right)\beta\sqrt{H^7S^3AK}\right)$$

Proof. We first rewrite the expression:

$$\begin{aligned}&\mathbb{E}\left[\sum_{k,h}\left|\left\langle P_k(s_{k,h},a_{k,h})-\widehat{P}_k(s_{k,h},a_{k,h}),V_{k,h+1}\right\rangle\right|\right]\\ &\leq\mathbb{E}\left[\sum_{k,h}\|\epsilon^p_k(s_{k,h},a_{k,h})\|_1\|V_{k,h+1}\|_\infty\right]\end{aligned}$$

where $\epsilon^p_k(s_{k,h},a_{k,h}) = P_k(s_{k,h},a_{k,h})-\widehat{P}_k(s_{k,h},a_{k,h})$. Using Cauchy-Schwarz inequality, we rewrite this as:

$$\sqrt{\mathbb{E}\left[\sum_{k,h}\|\epsilon^p_k(s_{k,h},a_{k,h})\|_1^2\right]}\sqrt{\mathbb{E}\left[\sum_{k,h}\|V_{k,h+1}\|^2_\infty\right]}$$

For bounding the sum of values under the second square root, we can directly use the Lemma 8 from Russo (2019):

$$\sqrt{\mathbb{E}\left[\sum_{k,h}\|V_{k,h+1}\|^2_\infty\right]} = \widetilde{\mathcal{O}}(H^3\sqrt{SK}) \tag{32}$$

For bounding the expected estimation error, we consider two events: $F^{(p)}$ when the confidence widths are incorrect and $(F^{(p)})^c$ when the confidence intervals are valid for all (s,a), k and h. Therefore, we have:

$$\begin{aligned}&\mathbb{E}\left[\sum_{k,h}\|\epsilon^p_k(s_{k,h},a_{k,h})\|_1^2\right]\\ &=\mathbb{E}\left[\sum_{k,h}\|\epsilon^p_k(s_{k,h},a_{k,h})\|_1^2|F^{(p)}\right]P(F^{(p)})\\ &\quad+\mathbb{E}\left[\sum_{k,h}\|\epsilon^p_k(s_{k,h},a_{k,h})\|_1^2|(F^{(p)})^c\right]P((F^{(p)})^c)\end{aligned}$$

Setting $\delta_p = 1/KH$, we can bound the sum under failure event to a constant. For the other term, we see that it is

equivalent to:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k,h}\|\epsilon_k^p(s_{k,h},a_{k,h})\|_1^2|(F^{(p)})^c\right]P((F^{(p)})^c)\\
&\leq \mathbb{E}\left[\sum_{k,h}\left(1\wedge\beta\sqrt{S\gamma_k(s_{k,h},a_{k,h})}\|x_k\|_{Z^{-1}_{k,sa,h}}\right)^2\right]\\
&\leq \frac{2\beta^2 S\bar{\gamma}_K}{\eta\alpha}\mathbb{E}\left[\sum_{k,h}\left(1\wedge\frac{\eta\alpha}{2}\|x_k\|^2_{Z^{-1}_{k,sa,h}}\right)\right]\\
&= \widetilde{\mathcal{O}}\left(\left(\frac{d\max_{s,a}\|W^*_{sa}\|_F^2}{\alpha}+\frac{d^2}{\alpha^2}\right)\beta^2S^2AH\right) \quad (33)
\end{aligned}$$

Combining eqs. (32) and (33), we get the desired result. □

The final regret guarantee can be obtained by adding terms from Lemma B.9 and Lemma B.10.

C PROOF OF MISTAKE BOUND FROM SECTION 4.1.2

In order to prove the mistake bound, we need to bound the number of episodes where the policy's value is more than ϵ-suboptimal. We start with inequality (24):

$$\begin{aligned}
&V^*_{k,1}(s)-V^{\pi_k}_{k,1}(s)\\
&\leq \sum_{h,s,a}\mathbb{P}_k[s_h,a_h=s,a|s_{k,1}](2\varphi_{k,h}(s,a)\wedge V_h^{\max})
\end{aligned}$$

We note that if $\varphi_{k,h}(s,a)\leq\frac{\epsilon}{2H}$ for all k,h and (s,a), then we have

$$\begin{aligned}
&V^*_{k,1}(s)-V^{\pi_k}_{k,1}(s)\\
&\leq \sum_{h,s,a}\mathbb{P}_k[s_h,a_h=s,a|s_{k,1}]\frac{\epsilon}{H}\\
&\leq \epsilon
\end{aligned}$$

In order to satisfy the constraint, we bound each error term as: $\xi^{(p)}\leq\frac{\epsilon}{4H^2}$ and $\xi^{(r)}\leq\frac{\epsilon}{4H}$.

We bound the number of episodes where this constraint is violated. For simplicity, we consider that the rewards are known and only consider the transition probabilities in the analysis:

$$\begin{aligned}
&\sum_{k\in[K]}\mathbb{I}\left[\exists(s,a)\text{ s.t. }\xi^{(p)}_{k,sa}\geq\frac{\epsilon}{4H^2}\right]\\
&\leq\sum_{k\in[K]}\sum_{s,a}\mathbb{I}\left[\beta\sqrt{S}\sqrt{\gamma_{k,sa}}\|x_k\|_{Z^{-1}_{k,sa}}\geq\frac{\epsilon}{4H^2}\right]\\
&\leq\sum_{k\in[K]}\sum_{s,a}\frac{16\beta^2SH^4\gamma_{k,sa}}{\epsilon^2}\|x_k\|^2_{Z^{-1}_{k,sa}} \quad (34)\\
&\leq\frac{16\beta^2SH^4\gamma_{K+1}}{\epsilon^2}\sum_{k\in[K]}\sum_{s,a}\|x_k\|^2_{Z^{-1}_{k,sa}}\\
&\leq\frac{16\beta^2H^4\gamma_{K+1}}{\epsilon^2}\sum_{s,a}\sum_{k\in[K]}\|x_k\|^2_{Z^{-1}_{k,sa}} \quad (35)
\end{aligned}$$

where in the intermediate steps, we have used the nature of the indicator function and the fact that minimum is upper bounded by the average. Assuming that $N_{k,sa}$ denotes the number of visits to pair (s,a) in episode k, we rewrite the inner term as:

$$\begin{aligned}
\|x_k\|^2_{Z^{-1}_{k+1,sa}}&=x_k^\top(Z_k+N_{k,sa}x_kx_k^\top)^{-1}x_k\\
&=x_k^\top Z_{k,sa}x_k-\frac{N_{k,sa}x_k^\top Z^{-1}_{k,sa}x_kx_k^\top Z^{-1}_{k,sa}x_k}{1+N_{k,sa}x_k^\top Z^{-1}_{k,sa}x_k}\\
&=\|x_k\|^2_{Z^{-1}_{k,sa}}-\frac{N_k\|x_k\|^4_{Z^{-1}_{k,sa}}}{1+N_{k,sa}\|x_k\|^2_{Z^{-1}_{k,sa}}}
\end{aligned}$$

With this setup, we get:

$$\begin{aligned}
\|x_k\|^2_{Z^{-1}_{k,sa}}&=\frac{\|x_k\|^2_{Z^{-1}_{k+1,sa}}}{1-N_{k,sa}\|x_k\|^2_{Z^{-1}_{k+1,sa}}}\\
&\leq\frac{\lambda+H}{\lambda}\|x_k\|^2_{Z^{-1}_{k+1,sa}}\\
&\leq\frac{\lambda+H}{\lambda}\langle Z^{-1}_{k+1,sa},N_{k,sa}x_kx_k^\top\rangle
\end{aligned}$$

Using Lemma 11 from Hazan et al. (2007), the inner sum in eq. (35), can be bounded as:

$$\frac{\lambda+H}{\lambda}\sum_{k\in[K]}\|x_k\|^2_{Z^{-1}_{k+1}}\leq d\log\left(\frac{R^2KH}{\lambda}+1\right)$$

Combining all these bounds, we get:

$$\begin{aligned}
&\sum_{k\in[K]}\mathbb{I}\left[\exists(s,a)\text{ s.t. }\xi^{(p)}_{k,sa}\geq\frac{\epsilon}{4H^2}\right]\\
&\leq\frac{16(\lambda+H)\beta^2dS^2AH^4\gamma_{K+1}}{\lambda\epsilon^2}\log\left(\frac{R^2KH}{\lambda}+1\right)
\end{aligned}$$

Noting that $\gamma_{K+1}=\mathcal{O}\left(\frac{d\log^2KH}{\alpha}+S\right)$, we get the final mistake bound as:

$$\mathcal{O}\left(\frac{dS^2AH^5\log KH}{\epsilon^2}\left(\frac{d\log^2KH}{\alpha}+S\right)\right)$$

ignoring $\mathcal{O}(\text{poly}(\log\log KH))$ terms.

D PROOF OF THE LOWER BOUND

Proof. We start with the lower bound from Jaksch et al. (2010) adapted to the episodic setting.

Theorem D.1 (Jaksch et al. (2010), Thm. 5)**.** *For any algorithm $\mathbf{A}'$, there exists an MDP M with S states, A actions, and horizon H, such that for $K \geq dSA$, the expected regret of $\mathbf{A}$ after K episodes is:*

$$\mathbb{E}[R(K; \mathbf{A}', s, M)] = \Omega(H\sqrt{SAK})$$

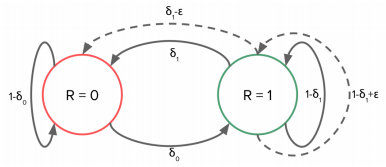

Figure 1: Hard 2-state MDP (Osband and Van Roy, 2016)

The lower bound construction is obtained by concatenating $\lceil S/2 \rceil$-copies of a bandit-like 2-state MDP as shown in figure 1[9]. Essentially, state 1 is a rewarding state and all but one action take the agent to state 0 with probability δ_1. The remaining optimal action transits to state 0 with probability $\delta_1 - \epsilon$. This makes the construction similar to a hard Bernoulli multi-armed bandit instance which leads to the lower bound. Now, we will construct a set of such hard instances with the logit link function for transition probabilities. A similar construction for the linear combination case is discussed in Appendix D. Since, the number of next states is 2, we use a GLM with parameter vector w^* of shape $1 \times d$. Thus, for any context x, the next state probabilities are given as:

$$p(1|1, a; x) = \frac{\exp(w_a^* x)}{1 + \exp(w_a^* x)} = \phi(w_a^* x)$$

If $w_a^* x = 0$, the value turns out to be $\frac{1}{2}$ which we choose as $\delta_1 - \epsilon$. For making the probability $\delta_1 = \frac{1}{2} + \epsilon$, we need to have $w_a^* x = \phi^{-1}(\delta_1) = c^*$. We consider the case where for each index i, all but one action has $w_a^*[i] = 0$ and one action a_i^* has $w_{a^*}^*[i] = c^*$. The sequence of contexts given to the algorithm comprises of K/d indicator vectors with 1 at only one index. Therefore, for each episode k, we get an MDP with $p_k(0|1, a^*_{k\%d}) = 1/2$ for one optimal action and $1/2$ for all other actions. Therefore, this is a hard instance as shown in figure 1. The agent interacts with each such MDP $K_i \approx K/d$ times. Further, these MDPs are decoupled as the context vectors are non-overlapping. Therefore, we have:

$$\begin{aligned}
&\mathbb{E}[R(K; \mathbf{A}, M_{1:K}, s_{1:K})] \\
&= \sum_{i=1}^{d} \mathbb{E}[R(K_i; \mathbf{A}, M_{1:K}, s_{1:K})] \\
&\geq \sum_{i=1}^{d} cH\sqrt{SAK/d} = cH\sqrt{dSAK}
\end{aligned}$$

□

Linear combination case Similar to the logit case, we need to construct the sequence of hard instances in the linear combination case. It turns out that a similar construction works. Note that, in the linear combination case, each parameter vector w_a^* now directly contains the probability of moving to the rewarding state. In other words, each index of this vector $w_a^*[i]$ corresponds to the next state visitation probability for the base MDP M_i. Therefore, for each index, we again set one action's value to $\frac{1}{2} + \epsilon$ and all others to 0. This maintains the independence argument and using indicator vectors as contexts, we get the same sequence of MDPs. The same lower bound can therefore be obtained for the linear combination case.

E OMITTED PROOFS FROM SECTION 6

Theorem E.1 (Multinomial GLM Online-to-confidence set conversion)**.** *Assume that loss function l_i defined in eq. (5) is α-strongly convex with respect to Wx. If an online learning oracle takes in the sequence $\{x_i, y_i\}_{i=1}^t$, and produces outputs $\{W_i\}_{i=1}^t$ for an input sequence $\{x_i, y_i\}_{i=1}^t$, such that:*

$$\sum_{i=1}^{t} l_i(W_i) - l_i(W) \leq B_t \quad \forall W \in \mathcal{W}, t > 0,$$

then with $\overline{W}_t$ as defined above, with probability at least $1 - \delta$, for all $t \geq 1$, we have

$$\|W^* - \overline{W}_t\|^2_{Z_{t+1}} \leq \gamma_t$$

where $\gamma_t := \gamma'_t(B_t) + \lambda B^2 S - (\|C_t\|_F^2 - \langle \overline{W}_t, X_t^\top C_t \rangle)$,

$$\gamma'_t(B_t) := 1 + \tfrac{4}{\alpha} B_t + \tfrac{8}{\alpha^2} \log\left(\tfrac{1}{\delta}\sqrt{4 + \tfrac{8B_t}{\alpha} + \tfrac{16}{\alpha^4 \delta^2}}\right).$$

Proof. Using the strong convexity of the losses l_i, we again have:

$$\begin{aligned}
&l_i(W_i) - l_i(W^*) \\
&\geq \langle \nabla l_i(W^*), W^* - W_i \rangle + \frac{\alpha}{2} \|W^* x_i - W_i x_i\|_2^2
\end{aligned}$$

[9] The two state MDP is built using $A/2$ actions with the rest used for concatenation. We ignore this as it only leads to a difference in constants.

Summing this for $i = 1$ to t and substituting the regret bound B_t, we get

$$\sum_{i=1}^{t} \|W^* x_i - W_i x_i\|_2^2 \leq \frac{2}{\alpha} B_t + \frac{2}{\alpha} \sum_{i=1}^{t} \langle p_t - y_t, W^* x_i - W_i x_i \rangle \tag{36}$$

Now, we focus on bounding the second term in the `rhs`. We note that for any $z \in \mathbb{R}^S$, we have

$$\langle p_t - y_t, z \rangle \leq \|p_t - y_t\|_2 \|z\|_2 \leq 2\|z\|_2$$

In addition, $\langle \eta_t, z \rangle \coloneqq \langle p_t - y_t, z \rangle$ is a martingale with respect to the filtration $\mathcal{F}_t \coloneqq \sigma(x_1, y_1, \ldots, x_{t-1}, y_{t-1}, x_t)$. This shows that

$$\mathbb{E}[D_t^\lambda | \mathcal{F}_t] = \mathbb{E}[\exp(\lambda \langle \eta_t, z \rangle - \tfrac{1}{2}\lambda^2 \|z\|_2^2) | \mathcal{F}_t] \leq 1 \tag{37}$$

We can substitute $z_t = W^* x_t - W_t x_t$ which is $\mathcal{F}_t$ measurable. Now, using $S_t = \sum_{i=1}^{t} \langle \eta_i, z_i \rangle$, ineq. (37) implies that $M_t^\lambda = \exp\left(4\lambda S_t - \frac{1}{2}\lambda^2 \sum_{i=1}^{t} \|z_i\|_2^2\right)$ is a $\mathcal{F}_{t+1}$-adapted supermartingale. Using the same analysis as in Abbasi-Yadkori et al. (2012), we get the following result:

Corollary E.2 (Corollary 8, Abbasi-Yadkori et al. (2012))**.** *With probability at least $1 - \delta$, for all $t > 0$, we have*

$$\sum_{i=1}^{t} \langle \eta_i, z_i \rangle \leq \sqrt{2\left(1 + \sum_{i=1}^{t} \|z_i\|_2^2\right) \ln\left(\tfrac{1}{\delta}\sqrt{(1 + \sum_{i=1}^{t} \|z_i\|_2^2)}\right)}$$

Substituting this in ineq. (36), we get

$$\sum_{i=1}^{t} \|z_i\|_2^2 - \frac{2}{\alpha} B_t \leq \frac{2}{\alpha}\sqrt{2\left(1 + \sum_{i=1}^{t} \|z_i\|_2^2\right) \ln\left(\tfrac{1}{\delta}\sqrt{(1 + \sum_{i=1}^{t} \|z_i\|_2^2)}\right)}$$

We now use Lemma 2 from Jun et al. (2017), to obtain a simplified bound:

Lemma E.3 (Lemma 2, Jun et al. (2017))**.** *For $\delta \in (0, 1)$, $a \geq 0, f \geq 0, q \geq 1$, $q^2 \leq a + fq\sqrt{\log \frac{q}{\delta}}$ implies*

$$q^2 \leq 2a + f^2 \log\left(\frac{\sqrt{4a + f^4/(4\delta^2)}}{\delta}\right)$$

With $q \coloneqq \sqrt{1 + \sum_{i=1}^{t} \|z_i\|_2^2}$, $a \coloneqq 1 + \frac{2}{\alpha} B_t$ and $f = \frac{2\sqrt{2}}{\alpha}$, we now have:

$$\sum_{i=1}^{t} \|W^* x_i - W_i x_i\|_2^2 \leq \gamma_t' \tag{38}$$

with $\gamma_t' \coloneqq 1 + \frac{4}{\alpha} B_t + \frac{8}{\alpha^2} \log\left(\frac{1}{\delta}\sqrt{4 + \frac{8B_t}{\alpha} + \frac{16}{\alpha^4 \delta^2}}\right)$.

We can rewrite ineq. (38) as

$$\|X_t W^{*\top} - C_t\|_F^2 \leq \gamma_t' \tag{39}$$

If we center this quadratic form around

$$\begin{aligned} \overline{W}_t &\coloneqq \arg\min_W \|X_t W^\top - C_t\|_F^2 + \lambda \|W\|_F^2 \\ &= Z_{t+1}^{-1} X_t^\top C_t \end{aligned}$$

we can rewrite the set as:

$$\begin{aligned} &\|W^* - \overline{W}_t\|_{Z_{t+1}}^2 \\ &\leq \lambda B_p^2 S + \gamma_t' - \left(\|\overline{W}_t\|_F^2 + \|X_t \overline{W}_t^\top - C_t\|_F^2\right) \end{aligned}$$

Simplifying the expression on the `rhs` gives the stated result. □